Методы коррекции локально возмущенных полуметрик

Громов И. А.
igor_gromov@mail.ru
Москва, МГУ им. М. В. Ломоносова

Для решения задач интеллектуального анализа данных широко применяются метрические методы. Эффективность их использования существенно зависит от выбора функции сходства (например, полуметрики) на обрабатываемых объектах. Нередко выбор полуметрики для решения конкретной задачи субъективен, а значит, можно модифицировать полуметрику с тем, чтобы она точнее отражала характерные сходства и различия исследуемых объектов.

Постановка задачи. Рассматривается задача коррекции полуметрики в ситуации, когда эксперт (в предметной области) требует изменить расстояние между ровно одной парой объектов. Пусть дано конечное множество объектов мощности N с заданной на нем полуметрикой R. Эксперт выбирает одну пару объектов (i, j) и по своему усмотрению изменяет расстояние между ними: $r_{ij} \mapsto r'_{ij}$. В общем случае это влечет нарушение неравенств треугольника. Требуется предложить методы коррекции $A\colon R' \mapsto \widetilde{R}$, которые позволят построить полуметрику $\widetilde{R}$ такую, что $r'_{ij} = \tilde{r}_{ij}$, и при этом будут

1) универсальны, т. е. применимы для любых R и R';
2) «наиболее полно» соответствовать экспертной интерпретации внесенного возмущения;
3) строить полуметрику $\widetilde{R}$, «максимально схожую» с исходной полуметрикой R.

Формализация задачи. Для того, чтобы коррекция отвечала требованию 2, предложена процедура обучения алгоритма коррекции. Она позволяет взаимодействовать с экспертом на более естественном для него языке и полнее учитывать его требования к проводимой коррекции.

Для формализации понятия сходства метрик (а при некоторых допущениях также и полуметрик) введен ряд интерпретируемых функционалов их сравнения:

$$Q_w(R, \widetilde{R}) = w_u Q_u(R, \widetilde{R}) + w_p Q_p(R, \widetilde{R}), \qquad w_u, w_p \geqslant 0; \tag{1}$$

$$Q_u(R, \widetilde{R}) = \frac{\sum\limits_{(kl)\in E_N} w_{kl} \left(\tilde{r}_{kl} - r_{kl}\right)^2}{\sum\limits_{(kl)\in E_N} \tilde{r}_{kl}^2}, \qquad w_{kl} \geqslant 0; \tag{2}$$

$$Q_p(R,\widetilde{R}) = \sum_{\triangle(klm)} w_{kl,km}\left(\frac{\tilde{r}_{kl}}{\tilde{r}_{km}} - \frac{r_{kl}}{r_{km}}\right)^2 + \\ + w_{kl,lm}\left(\frac{\tilde{r}_{kl}}{\tilde{r}_{lm}} - \frac{r_{kl}}{r_{lm}}\right)^2 + w_{km,lm}\left(\frac{\tilde{r}_{km}}{\tilde{r}_{lm}} - \frac{r_{km}}{r_{lm}}\right)^2, \qquad (3)$$
$$w_{kl,km},\ w_{kl,lm},\ w_{km,lm} \geqslant 0.$$

В ряде случаев (для некоторых значений весов в (2), (3)) получены методы коррекции, доставляющие минимум указанным функционалам. Таким образом, удалось напрямую согласовать функционалы сравнения метрик с формулами коррекции.

Трехэтапная схема построения алгоритмов коррекции полуметрики. Существуют различные подходы к решению поставленной задачи коррекции полуметрики. Автором предложена трехэтапная схема коррекции возмущенных полуметрик. В рамках данной схемы в ходе коррекции рассматриваются тройки попарно различных объектов и треугольники, в которых вершинами являются такие объекты, а длинами сторон — расстояния между ними. На первом этапе коррекции рассматриваются только тройки объектов вида (ijk), на втором — (ikl) и (jkl), на третьем — (klm), $k,l,m \notin \{i,j\}$. На каждом из этапов коррекции модифицируются расстояния в тех и только тех треугольниках, в которых неравенства треугольника нарушены. В общем случае данная схема коррекции требует исследования всех троек объектов и имеет сложность $O(N^3)$. Для достаточно больших значений N вычислительная сложность данной схемы становится препятствием к ее практическому применению.

В результате исследования конечных полуметрик были выявлены новые свойства. На их основании предложен ряд алгоритмов, не требующих рассмотрения в процессе коррекции всех троек объектов. Было доказано, что при их использовании достаточно проведения только первых двух этапов коррекции для построения полуметрики. Таким образом, данные алгоритмы имеют квадратичную $O(N^2)$, а в специальном случае — линейную сложность, т. е. полуметрика строится в ходе первого этапа. Вычислительная эффективность предложенных алгоритмов позволяет проводить преобразование полуметрики интерактивно.

Универсальный алгоритм коррекции полуметрики $\mathcal{A}$. В данном разделе сформулирован *универсальный* алгоритм коррекции полуметрики $\mathcal{A}$, т. е. алгоритм, гарантировано строящий полуметрику для любых исходных полуметрик и любых локальных возмущений, внесенных экспертом.

Пусть эксперт модифицировал в полуметрике R одно расстояние: $r_{ij} \mapsto r'_{ij}$ и требует сохранить указанное им значение r'_{ij}. Кроме того,

эксперт зафиксировал функционал сравнения полуметрик (тем самым давая интерпретацию внесенного возмущения). Тогда для того, чтобы скорректировать возникшие вследствие этого нарушения неравенств треугольника в R', предлагается следующий алгоритм $\mathcal{A}$.

1-й этап: коррекция $\triangle(ijk)$, в которых неравенства треугольника нарушены. Какой именно метод коррекции при этом должен быть применен, определяется выбором функционала сравнения полуметрик.

2-й этап: коррекция $\triangle(ikl)$ и $\triangle(jkl)$, в которых неравенства треугольника нарушены. Коррекция проводится по следующему правилу:

$$\tilde{r}_{kl} = \alpha \tilde{r}_{kl}^{\min} + (1-\alpha)\tilde{r}_{kl}^{\max},\ \forall (k,l) : k,l \notin \{i,j\},$$

где $\tilde{r}_{kl}^{\min} = \max\{|\tilde{r}_{ik} - \tilde{r}_{il}|, |\tilde{r}_{jk} - \tilde{r}_{jl}|\}$, $\tilde{r}_{kl}^{\max} = \min\{(\tilde{r}_{ik} + \tilde{r}_{il}), (\tilde{r}_{jk} + \tilde{r}_{jl})\}$, $\alpha \in [0,1]$ и α фиксировано *для всех* $\triangle(ikl)$, $\triangle(jkl)$.

3-й этап: не требуется.

На первом этапе выполнения алгоритма $\mathcal{A}$ требуется рассмотреть $N-2$ треугольников, на втором — $(N-2)(N-3)$ треугольников. Если сложность вычисления первого этапа — $O(N)$, то сложность алгоритма $\mathcal{A}$ — $O(N^2)$.

Параметр α может быть либо задан экспертом, либо настроен для получения величины $\tilde{r}_{kl}$, наиболее близкой к r_{kl}.

В универсальном алгоритме $\mathcal{A}$ не накладывается никаких специальных ограничений на методы коррекции, используемые на первом этапе, что, однако, компенсируется весьма жесткими условиями, налагаемыми на процедуру коррекции на втором. В ходе исследования алгоритмов коррекции в рамках трехэтапной схемы удалось перераспределить мощность требований между этапами выполнения алгоритма в духе алгебраического подхода к синтезу алгоритмов распознавания. За счет сужения множества методов коррекции на первом этапе были существенно смягчены условия второго этапа.

Литература

[1] *Майсурадзе А. И.* Гомогенные и ранговые базисы в пространствах метрических конфигураций // ЖВМиМФ. — 2006. — Т. 46, № 2. — С. 344–361.

[2] *Deza M., Dutour M.* Data mining for cones of metrics, quasi-metrics, semimetrics, and super-metrics. — 2006.