Motivation

Aggregation

Mergeable summaries

Questions

References

Complex aggregation of large data sets

Vladimir Batagelj

IMFM Ljubljana, IAM UP Koper, NRU HSE Moscow

7 th Symbolic Data Analysis workshop 2018

Viana do Castelo, 18-20. October 2018

Outline

1 Motivation
2 Aggregation
3 Mergeable summaries
4 Questions

Vladimir Batagelj: vladimir.batagelj@fmf.uni-lj.si

Last version of slides (October 18, 2018, 23 : 40): SDAggreg.pdf

Motivation

For the representation of symbolic data by discrete distributions $(n, \mathbf{p})$ used in our program Clamix (Batagelj et al., 2015) for clustering symbolic data we can observe two important properties

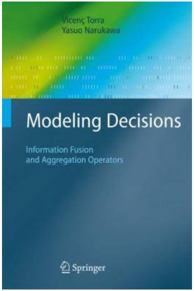
- fixed space required for a description of a unit/cluster;
- description of a union of two disjoint clusters can be obtained from their descriptions.

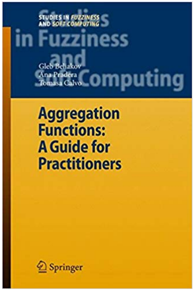
In this talk I will elaborate on this second observation.

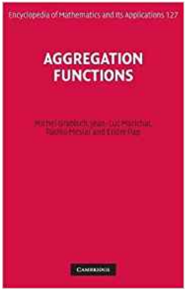
How to join the population pyramids for China and Albania?

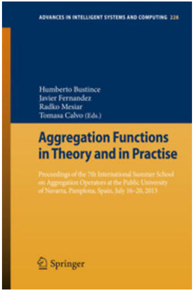
Aggregation

In analysis of large data sets the *aggregation* is a standard way for reducing size (complexity) of the data. Recently some books dealing with theoretical and algorithmic background of the traditional aggregation (replacing values of variable over a group by a single value) were published (Beliakov et al., 2007; Torra and Narukawa, 2007; Grabisch et al., 2009; Bustince et al., 2013).









Aggregation

Data analysis programs provide aggregation functions such as: means (arit, geom, harm, median, modus), min, max, product, bounded sum, counting, etc. A special care has to be given to variables measured in different measurement scales.

Aggregation

In theoretical discussion the traditional aggregation functions are usually "normalized" to the interval $[0, 1]$ – they take real arguments in $[0, 1]^k$ and produce a value in $[0, 1]$, and satisfy the conditions: $f(\mathbf{0}) = 0$, $f(\mathbf{1}) = 1$, and monotonicity $\mathbf{x} \leq \mathbf{y} \Rightarrow f(\mathbf{x}) \leq f(\mathbf{y})$. Often, in applications, also idempotency and symmetry are required.

The applications of traditional aggregation functions are, besides determining a representative value for a group of measurements, mainly to combine partial criteria into single criterion (multicriteria optimization and decision making) or to express the membership degree in combined fuzzy sets.

Aggregation

A problem with the traditional aggregation is that often too much of information is discarded thus reducing the precission of the obtained results.

A much better, preserving more information, summarization of original data can be achieved by representing aggregated data using selected types of complex data such us symbolic objects (Diday, 1988), compositions (Aitchison, 1986), functional data (Ramsay and Silverman, 2005), etc. In the SDA framework much work is devoted to the summarization process, for example the function `classic.to.sym` in RSDA (Rodriguez, 2018), and SODAS or SYR software.

Mergeable summaries

In complex data analysis the measured values over a selected group A are aggregated into a complex object $\Sigma(A)$ and not into a single value. Most of the theory does not apply directly.

In our contribution we present an attempt to start building a theoretical background of complex aggregation.

An interesting question is, which complex data types are compatible with merging of disjoint sets of units

$$\Sigma(A \cup B) = F(\Sigma(A), \Sigma(B)), \qquad \text{for} \quad A \cap B = \emptyset.$$

See also Diday (1995).

Mergeable summaries

Searching for a name I was inclined towards *hierarchical* or *mergeable summarization*. I recently tried this term on Google and surprise – *mergeable summaries* were proposed and elaborated by Agarwal et al. (2012).

They turn out to enable parallelization in big data algorithms and streams processing.

The summarization in big data is not deterministic and allows some error. A summary is *mergeable*, if error and space (size of summary) does not increase after the merge.

In my talk I will discuss *exactly mergeable* summaries “without errors”.

Exactly mergeable summaries

simple examples

We assume $A \cap B = \emptyset$

1. $\Sigma(A) = |A| = n_A$
   $\Sigma(A \cup B) = \Sigma(A) + \Sigma(B)$
2. $\Sigma(A) = \min(A)$
   $\Sigma(A \cup B) = \min(\Sigma(A), \Sigma(B))$
3. $\Sigma(A) = \max(A)$
   $\Sigma(A \cup B) = \max(\Sigma(A), \Sigma(B))$
4. $\Sigma(A) = (\text{First}(A), \text{Second}(A))$
   $\Sigma(A \cup B) = (\text{First}(L), \text{Second}(L))$, where
   $L = \{\text{First}(A), \text{Second}(A), \text{First}(B), \text{Second}(B)\}$
5. $\Sigma(A) = (n_A, \mu_A)$
   $\Sigma(A \cup B) = (n_A + n_B, \frac{n_A\mu_A + n_B\mu_B}{n_A + n_B})$
6. $\Sigma(A) = \sum_{X \in A} v(X)$
   $\Sigma(A \cup B) = \Sigma(A) + \Sigma(B)$

Exactly mergeable summaries

moments

For example, in physics and engineering the measurements are usually aggregated as $\mu \pm \sigma$. It would be better for measurements A to represent them as $\Sigma(A) = (n_A, \mu_A, \sigma_A)$, where n_A is the number of measurements.

Then additional measurements B, $A \cap B = \emptyset$, $\Sigma(B) = (n_B, \mu_B, \sigma_B)$ can be combined into measurements $C = A \cup B$, $\Sigma(C) = (n_C, \mu_C, \sigma_C)$ determined by $\Sigma(A)$ and $\Sigma(B)$ as follows

$$n_C = n_{A \cup B} = n_A + n_B$$

$$\mu_C = \mu_{A \cup B} = \frac{n_A \mu_A + n_B \mu_B}{n_C}$$

$$\sigma_C = \sigma_{A \cup B} = \sqrt{\frac{S_C}{n_C} - \mu_C^2}$$

where $S_C = S_A + S_B$ and $S_X = n_X(\sigma_X^2 + \mu_X^2)$.

The result can be extended to higher moments.

Exactly mergeable summaries

set membership count

Counting number of values from C in A

$$n(A; C) = |A \cap C|$$

is an exactly mergeable summary.

Proof::

$$n(A \cup B; C) = |(A \cup B) \cap C| = |(A \cap C) \cup (B \cap C)| =$$

$$= |A \cap C| + |B \cap C| - |A \cap B \cap C| = n(A; C) + n(B; C)$$

Combining exactly mergeable summaries

Let Σ_1 and Σ_2 be exactly mergeable summaries. Then also

$$\Sigma_1 \oplus \Sigma_2(A) = (\Sigma_1(A), \Sigma_2(A))$$

is an exactly mergeable summary.

Proof: $\Sigma_1 \oplus \Sigma_2(A \cup B) = (\Sigma_1(A \cup B), \Sigma_2(A \cup B)) =$

$$= (F_1(\Sigma_1(A), \Sigma_1(B)), F_2(\Sigma_2(A), \Sigma_2(B)))$$

Therefore, since set membership counts are exactly mergeable, the *barcharts*

$$C = \{X : v(X) = c\}$$

and *histograms*

$$C = \{X : v(X) \in [a, b)\}$$

are exactly mergeable summaries.

Proving that a summary is not exactly mergeable

If for a summary Σ exist sets A_1, B_1, A_2, B_2 such that $A_1 \cap B_1 = \emptyset$, $A_2 \cap B_2 = \emptyset$, $\Sigma(A_1) = \Sigma(A_2)$, $\Sigma(B_1) = \Sigma(B_2)$, and $\Sigma(A_1 \cup B_1) \neq \Sigma(A_2 \cup B_2)$ then Σ **is not** exactly mergeable.

Proof: Assume that Σ is exactly mergeable. Then

$$\Sigma(A_1 \cup B_1) = F(\Sigma(A_1), \Sigma(B_1)) = F(\Sigma(A_2), \Sigma(B_2)) = \Sigma(A_2 \cup B_2)$$

a contradiction.

Median is not exactly mergeable summary

$$\text{med}(A) = \text{order}(A)[\left\lceil \frac{n_A}{2} \right\rceil]$$

$A_1 = [3, 4, 1] \qquad \text{med}(A_1) = 3$
$B_1 = [9, 6] \qquad \text{med}(B_1) = 6$
$\text{med}(A_1 \cup B_1) = 5$

$A_2 = [3, 4] \qquad \text{med}(A_2) = 3$
$B_2 = [6, 2, 7] \qquad \text{med}(B_2) = 6$
$\text{med}(A_2 \cup B_2) = 4$

Second is not exactly mergeable summary

$$A_1 = [1, 3, 5] \qquad \mathrm{Second}(A_1) = 3$$
$$B_1 = [2, 5, 6] \qquad \mathrm{Second}(B_1) = 5$$
$$\mathrm{Second}(A_1 \cup B_1) = 2$$

$$A_2 = [3, 3, 6] \qquad \mathrm{Second}(A_2) = 3$$
$$B_2 = [4, 5, 7] \qquad \mathrm{Second}(B_2) = 5$$
$$\mathrm{Second}(A_2 \cup B_2) = 3$$

Questions

- How to measure the improvement of precision of results obtained by SDA ? Are they really performing better than the traditional methods based on averages?
- How to consider the preserved variability in criterion functions?
- Develop symbolic methods for big data mergeable summaries.

References I

Agarwal, PK., Cormode, G., Huang, Z., Phillips, J., Wei, Z., Yi. K. (2012). Mergeable summaries. In Proceedings of the 31st ACM SIGMOD-SIGACT-SIGAI symposium on Principles of Database Systems (PODS '12), Markus Krötzsch (Ed.). ACM, New York, NY, USA, 23-34.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.* Chapman and Hall, London.

Batagelj, V., Kejžar, N., Korenjak-Černe, S. (2015). Clustering of Modal Valued Symbolic Data. https://arxiv.org/abs/1507.06683.

Beliakov, G., Pradera, A, Calvo, T. (2007). *Aggregation Functions.* Springer.

Bustince, H., Fernandez, J., Mesiar, R., Calvo, T. (eds.). (2013). *Aggregation Functions in Theory and in Practise.* Proceedings of the 7th International Summer School on Aggregation Operators at the Public University of Navarra, Pamplona, Spain, July 16-20, 2013. Advances in Intelligent Systems and Computing 228. Springer.

References II

Diday, E. (1988). The symbolic approach in clustering and related methods of data analysis: The basic choices. In *Classification and related methods of data anlysis.* (H.-H. Bock, ed.), 673–684. North Holland, Amsterdam.

Diday, E. (1995). Probabilist, possibilist and belief objects for knowledge analysis. Annals of Operations Research. 55, pp. 227–276.

Grabisch, M., Marichal, J.-L., Mesiar, R., Pap, E. (2009). *Aggregation Functions.* Encyclopedia of Mathematics and its Applications 127. Cambridge UP.

Ramsay, J.O., Silverman, B.W. (2005). *Functional Data Analysis.* 2nd edition. Springer-Verlag, New York.

Rodriguez, O.R. (2018). *RSDA 2.0.5: R to Symbolic Data Analysis.* https://cran.r-project.org/web/packages/RSDA/.

Torra, V., Narukawa, Y. (2007). *Modeling Decisions: Information Fusion and Aggregation Operators.* Cognitive Technologies. Springer.

Acknowledgments

This work was supported in part by the Slovenian Research Agency (research program P1-0294 and research projects J7-8279 and J1-9187).