Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 17 (2013) iii–viii

Procedia Computer Science

Contents

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 17 (2013) 1121 – 1124

Procedia Computer Science

Information Technology and Quantitative Management (ITQM2013)

Super-threshold Procedures and Their Application to the Search Problem

Fuad Aleskerov[a,b], Evgeny Mitichkin[c], Sergey Shvydun[a] *, Vyacheslav Yakuba[b]

[a]National Research University Higher School of Economics, 11 Pokrovsky boulevard, Moscow 109028, Russia
[b]Institute of Control Sciences of Russian Academy of Sciences, 65 Profsoyuznaya street, Moscow 117997, Russia
[c]University of Mannheim, Schloss, Mannheim 68131, Germany

Abstract

The paper examines a choice problem in case of large number of alternatives characterized by a set of criteria. Very often existing choice procedures cannot be used due to their computational complexity. In contrast, the approach proposed in this paper utilizes a set of easy to use techniques having complexity less than the quadratic. We apply super-threshold procedures sequentially in order to find the best alternatives. The application to learning to rank problem is also considered. Finally, two methods of defining thresholds and order of criteria in super-threshold procedures are studied. While the first method is based on the etalon criterion, the second method uses the distribution function. Both methods developed are mainly oriented on the choice of search results in search engines and capable of processing large amount of data in networks, but can also be applied to any system in which the alternatives are represented by a set of criteria.

Selection and peer-review under responsibility of the organizers of the 2013 International Conference on Information Technology and Quantitative Management

Keywords: Super-threshold Choice Procedures; Search Problem, Learning to Rank; Adaptive Filtering.

1. Introduction

Consider a finite set A of alternatives evaluated by n criteria, i.e. the vector of criterial values $(x_1,\dots,x_n)$ is assigned to each alternative x from A

$$x \in A \to (x_1, \dots, x_n).$$

The problem lies in constructing a transformation φ — the rule of aggregation over A — such that $\varphi: A \to R^1$.

There are a lot of choice procedures applicable to cases with relatively small number of alternatives and criteria. However, in case of thousands of alternatives and criteria, existing procedures cannot provide sufficient results in a reasonable time. We propose to use super-threshold procedures for the choice of alternatives.

Super-threshold procedures are based on the rule that chooses all alternatives for which the estimate criterion value is greater or equal than some assigned threshold [1]. The main idea of this approach is to consider the consistent application of super-threshold procedures to different criteria in order to find the best alternatives. Comparing to other procedures the computational complexity of super-threshold procedures is lower (normally lower than quadratic) as every iteration decreases the number of alternatives.

The choice problem can be divided into two types. In the first type the information about the utility of alternatives is not specified. The utility of the alternative denotes how well this alternative meets the need of the end user. The utility of alternative can vary: the alternative with a high utility value is considered as the best alternative (choice) while alternative with a low utility value is considered as a bad alternative. In this case, the threshold values and the order of elimination are normally defined manually.

* Corresponding author. Tel.: +7-915-218-6094.
E-mail address: shvydun@hse.ru.

Selection and peer-review under responsibility of the organizers of the 2013 International Conference on Information Technology and Quantitative Management
doi:10.1016/j.procs.2013.05.143

The second type is the learning to rank problem which is based on the automatic construction of choice models using training dataset. The choice model is a set of rules that is used to choose the best alternatives. The training data consists of lists of items with partial order specified between items in each list. This order is typically induced by a numerical or ordinal score or a binary judgment (e.g. "relevant" or "not relevant") for each item. Choice model's purpose is to choose (select) items in a new, unseen list in a way, which is similar to the choice in the training data in some sense. In this case the threshold level as well as the order of procedures should be defined automatically.

In this paper, we propose two methods of how the order of criteria and the threshold value for each criterion can be defined. While the first method is based on the use of etalon criterion, the second method is based on the distribution function. Both methods developed are mainly oriented on the choice of search results in search engines and capable of processing large amount of data in networks, but can also be applied to any system in which the objects are represented by a set of criteria.

2. Super-threshold procedures

The main principle of the super-threshold procedures incorporates the sequential application of filters to the data, in order to build a system of constraints which is capable to automatically distinguish alternatives with a high utility values and other alternatives according to their criterial values.

Two methods of defining the order of elimination and threshold values are proposed

- Super-threshold etalon-based procedures,
- Super-threshold procedures based on the distribution function.

2.1. Super-threshold etalon-based procedures

All alternatives can be represented as points in (n+1)-dimensional space where n dimensions are criterial values of alternatives and one dimension is the utility. The procedure is divided into two steps. The first step of the procedure is the selection of criterion with the densest distribution of high-utility alternatives, i.e. the selection of criterion which criterion values of high-utility alternatives are closest to each other.

$$\varphi^* = \{\varphi_i \epsilon \varphi \mid \forall j, j \neq i: \max_{x \in A^*} \varphi_i(x) - \min_{x \in A^*} \varphi_i(x) \leq \max_{x \in A^*} \varphi_j(x) - \min_{x \in A^*} \varphi_j(x)\},$$

where φ is the set of criteria, A^* is the set of the best (high-utility) alternatives, $A^* \subset A$.

The chosen criterion is considered as the etalon. The direction of the super-threshold elimination depends on whether it is necessary to minimize or maximize the criterion. The threshold value can be either preset initially (mean, median, or other fixed value) or can be chosen or adjusted automatically.

The second step consists of construction of the order of the eliminations in order to increase the percentage of high-utility alternatives on the etalon criterion.

The order of elimination can be accomplished according to the following rule: the super-threshold choice is performed on the criterion (except the etalon) that excludes as much bad alternatives as possible within the accumulation (interval) of the best alternatives on the etalon criterion. When the criterion for elimination is found, it is included to the order of elimination and excluded from further consideration; alternatives, which values are below the threshold, are excluded from the set A and the procedure (recalculation the threshold and the order steps) iterates again. Thus, the procedure continues until the goal is achieved (only alternatives with a high utility value remain) or all criteria are used.

2.2. Super-threshold procedures based on the distribution function

All alternatives are divided into two groups – alternatives with high utility value and alternatives with low utility value (all other alternatives). Each group is presented by a set of distribution functions based on all criteria. In this case the distribution function describes the ratio of alternatives A for which the criterion value is less than or equal to value x.

The threshold value for each criterion is defined as the criterion value where the difference V* between two

distribution functions reaches its maximum.

$$V^* = \max_{x \in A} |F_1(x) - F_2(x)|,$$

where $F_1(x)$, $F_2(x)$ are the values of distribution functions depending on x.

It can be inferred that the concentration of alternatives with a high utility value among the others is highest at the point where the criterion value equals to x.

The order of elimination can be accomplished according to the following rule: the super-threshold choice is performed on the criterion with the highest difference V* between distribution function of high-utility alternatives and the distribution function of other alternatives. When the criterion for elimination is found, it is included to the order of elimination and excluded from further consideration; alternatives, the values of which are below the assigned threshold, are excluded and the procedure (recalculation of the threshold etc.) iterates again. As before, the procedure continues until the goal is achieved or all criteria are used.

As in etalon-based procedure, the result of the procedure based on the distribution function is the sequence of criteria accompanied with threshold values for the corresponding criteria (choice rule) that is used to define or predict the utility of new alternatives.

3. Application of super-threshold procedures to the search problem

A set from the LETOR 4.0 [2] collection was chosen to test two above-described methods. LETOR 4.0 is a special dataset with 10000 of queries provided by Microsoft Research to cope with the problem of learning to rank. It contains a set of evaluated web pages (documents). Each document is characterized by a set of criteria and has a relevance value. In information science and information retrieval, the relevance denotes how well a retrieved document or set of documents meets the information need of the user. The relevance judgments are obtained from a set of a commercial web search engine (Microsoft Bing), which take five values from 0 (irrelevant) to 4 (perfectly relevant). A query-document pair is represented by a 136-dimensional feature vector.

After a series of preliminary experiments some important properties of the set were noted, namely:

- Different queries and sets of documents are represented by different key features
- Maximization or minimization algorithms applied to the whole dataset do not choose (select) only the best documents with a high accuracy.
- It was noticed that simultaneous application of super-threshold procedures on several key (significant) features can lead to the required result.

Thus, the methods were tested on a subset of queries with the same choice rules. The testing was performed with the use of key (significant) features that were found by some procedures of data analysis such as the data transformation and the dimension reduction [3]. Key (significant) features are features that most affect the choice, so their existence is essential to the construction of the choice model. The main task of both methods was to estimate their average accuracy that can be defined as the ratio of the best alternatives in the remaining subset. In other words, the average accuracy shows how well both methods can distinguish relevant and irrelevant documents according to their feature's values. The average accuracy of the first method on the chosen set is 65%; the average accuracy of the second method is 46%. The first method showed better result on small dataset, the second demonstrated higher results in the case with different choice rules.

The computational complexity of both methods is also important. Both algorithms have the complexity that is less than quadratic, so the increase of the number of alternatives does not influence so much the computational complexity of both methods.

4. Conclusion

Super-threshold procedures are proposed to be used for the choice problem with large number of alternatives characterized by a set of criteria. One of the key advantages of the procedures is that the computational complexity is less than quadratic as every iteration decreases the number of alternatives. Two methods of defining the threshold and the elimination order in learning to rank problem are proposed. The first method is based on the use of the etalon criteria; the second method relies on distribution functions. Both procedures were tested on the LETOR 4.0

dataset. The experiments showed that the goal to develop self-learning choice model for the choice of alternatives is achieved.

References

1. M. Aizerman, F. Aleskerov F., Theory of Choice, Elsevier, North-Holland, 1995.

2. H. Li, T.Y. Liu, T. Qin, J. Xu, LETOR: A Benchmark Collection for Research on Learning to Rank for Information Retrieval, Information Retrieval Journal, Vol. 13, No. 4, 346-374, 2009.

3. E.M. Braverman, I.B. Muchnik, The Structural Methods for Empirical Data Processing, NAUKA Publishers, Moscow, 1983.