Maximizing Vector Distances for Purpose of Searching—A Study of Differential Evolution Suitability

Martin Kolařík, Roman Jašek, and Zuzana Komínková Oplatková

Tomas Bata University in Zlin, Faculty of Applied Informatics,
Nám. T.G. Masaryka 5555, 760 01 Zlín, Czech Republic
martin.kolarik@email.cz, {jasek,oplatkova}@fai.utb.cz

Abstract. This paper explores suitability of using of differential evolution for maximizing of weighted distances of vectors in a set of vectors. Increase in vector distances simplifies searching for the best matching vector what is a common task in many areas (for instance in biometric identification of people). Maximizing of weighted distances itself is complex and nonlinear problem. The differential evolution is efficient enough and helps in decreasing of the computational complexity space compared to enumerative methods where all possible combinations are calculated. To find out, if differential evolution can help with the problem, model experiments were introduced and executed. Experiments showed that differential evolution is able to resolve the problem.

Keywords: optimization, differential evolution, distance metric, nonlinear.

1 Introduction

Behaviometric identification systems [1] todays are based on values measuring various dynamic processes linked to human body (like gait or computer mouse movements). Measured values form a complex set of data where internal relations or dependencies are only hardly to determine. The fact is problematic because classification of data (and consequently of identified persons) requires as good understanding data as possible. Any technique which helps with decoding of meaning of intricate biometric data, or which improves orientation in the data, helps with the whole identification system and improves overall security of the system.

Organizing data into set of vectors is common, not only in identification systems. Usually each vector describes an entity (e.g. person) and components of vectors correspond to particular features of entities. To look up an entity in the set or to determine which entity is the closest to a given one whole set must be searched. Exact searching for an entity is easy, looking for the closest entity is more difficult. This paper explores possibility to improve searching for the best match by expanding distances between entities in the set. The idea is that among more distant entities it is easier to decide about where unknown entity belongs.

Maximizing of weighted distances in a set of vectors has analytic solution only for two-component vectors if analytical metric (e.g. arithmetic average of Euclidean distances) is used. If components n are more than 2, solution leads to nonlinear equation having $n-1$ unknown continuous variables (weights) from interval $[0..1]$. Because sum

P. Krömer et al. (eds.), *Proceedings of the Fifth Intern. Conf. on Innov. in Bio-Inspired Comput. and Appl. IBICA 2014,* Advances in Intelligent Systems and Computing 303,
DOI: 10.1007/978-3-319-08156-4_42,

of weights must be 1 and weights can mix arbitrarily, the resulting problem has nearly infinite complexity. If statistical metric is used (e.g. median), analytical solution does not exist at all. This is the first reason why we started to explore stochastic algorithm [2]. The second reason is that real data has unknown structure, frequently with hidden dependencies or duplicities between features, what adds yet more complexity. Among stochastic optimizing algorithms we decided to test differential evolution [3] (DE), because it is designed to work over vectors of real numbers. The authors have also previous experience with good performance of differential evolution which helps in decreasing of the computational complexity space compared to enumerative methods in general.

To explore suitability of differential evolution we proposed and executed more compound experiments: we tested performance of five statistical properties (arithmetic average, geometric average, summation, median and minimum) of two metrics of measuring distances of entities (Manhattan and Euclidean) [4], we combined and compared random variables of four distributions (normal, Weibull, Pearson and uniform) and we applied everything to data sets with two sizes.

Because we did not presume anything about DE's performace we proposed to use principal component analysis [5] (PCA) as independent validation method. We chose PCA, because expanding distances is linked to informational content of entities in some way—features with more variability (entropy, variance) should contribute to increasing of distances more than features with low variability—and PCA is a method for discovering features with the biggest variance. To detect if DE results can be validated using PCA we proposed and executed another experiment.

2 Theory

Before all, an overview of used terms and math is presented in a few chapters.

2.1 Vectors, Entities, Set of Vectors, Weight Vector

Vector is an ordered row of numbers:

$$a = \{a_1, a_2, ..., a_n\} = \{a_i\}$$

where a_i are components of the vector a and n is number of vector's components. In real case, vector corresponds to real entity (e.g. a person) and components of vectors correspond to real features (e.g. height of the person, his or her age etc.). Vector models the real object so it is possible to use names *entity* and *feature* instead of names *vector* and *component*.

Many vectors of the same size form together a set, *population* $\mathscr{P}$ of entities. Particular vector component represents always the same feature in all vectors:

$$\begin{aligned} e_1 &= \{f_{11}, f_{12}, ..., f_{1n}\} = \{f_{1i}\} \\ e_2 &= \{f_{21}, f_{22}, ..., f_{2n}\} = \{f_{2i}\} \\ &... \\ e_p &= \{f_{p1}, f_{p2}, ..., f_{pn}\} = \{f_{si}\} \end{aligned} \tag{1}$$