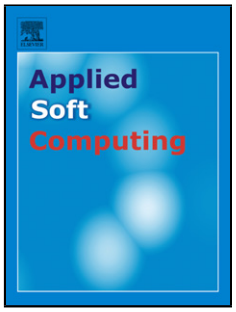

Title: A new evolutionary algorithm for mining top-*k* discriminative patterns in high dimensional data

Author: TarcÃ-sio Lucas TÃºlio C.P.B. Silva Renato Vimieiro Teresa B. Ludermir

PII:	S1568-4946(17)30313-7
DOI:	http://dx.doi.org/doi:10.1016/j.asoc.2017.05.048
Reference:	ASOC 4252
To appear in:	*Applied Soft Computing*
Received date:	30-12-2016
Revised date:	17-5-2017
Accepted date:	24-5-2017

A new evolutionary algorithm for mining top-*k* discriminative patterns in high dimensional data

Tarcísio Lucas[a,*], Túlio C. P. B. Silva[a], Renato Vimieiro[a], Teresa B. Ludermir[a]

[a]*Centro de Informática, Universidade Federal de Pernambuco,*
Av. Jornalista Anibal Fernandes, s/n - Cidade Universitária (Campus Recife)
50.740-560 - Recife - PE - BRAZIL

Abstract

This paper presents an evolutionary algorithm for Discriminative Pattern (DP) mining that focuses on high dimensional data sets. DPs aims to identify the sets of characteristics that better differentiate a target group from the others (e.g. successful vs. unsuccessful medical treatments). It becomes more natural to extract information from high dimensionality data sets with the increase in the volume of data stored in the world (30GB/s only in the internet). There are several evolutionary approaches for DP mining, but none focusing on high-dimensional data. We propose an evolutionary approach attributing features that reduce the cost of memory and processing in the context of high-dimensional data. The new algorithm thus seeks the best (top-*k*) patterns and hides from the user many common parameters in other evolutionary heuristics such as population size, mutation and crossover rates, and the number of evaluations. We carried out experiments with real-world high-dimensional and traditional low dimensional data. The results showed that the proposed algorithm was superior to other approaches of the literature in high-dimensional data sets and competitive in the traditional data sets.

Keywords: subgroup discovery, evolutionary algorithms, discriminative patterns, high dimensional data.

1. Introduction

This paper presents an evolutionary algorithm for Discriminative Pattern (DP) mining that focuses on high dimensional data sets. Discriminative pattern mining is a data mining task that has the objective of identifying sets of items that distinguish a target group from the others, for example: successful from unsuccessful treatments, unhealthy from healthy cells, spam from other emails, or even positive from negative sentiments in sentiment analysis. The necessity to investigate new methods, especially heuristics methods mining these patterns, comes from the fact that data generated/collected from many domains have different characteristics from those of last decade. The vast amount and high dimensionality of data sets in this so-called *Era of Big Data* render the application of existing methods infeasible.

Studies estimate that in the internet alone around 30GB of data, including texts, images and videos, are produced each second. Another important source of data is the biomedical sciences, particularly the *Omics* (genomics, proteomics, transcriptomics, ...), since the price for sequencing samples has dramatically dropped in the last years. These areas have two things in common: (1) they are major contributors to today's massive amount of available data (big data); (2) data from these domains is usually very high dimensional with tens of thousands to millions of attributes. In this sense they present new challenges to data mining and machine learning researchers. Among these challenges is the need for new tools for exploratory data analysis.

Discriminative patterns are an important tool for exploratory data analysis, since recurring patterns in the data are summarized in a simple way [1]. This is particularly suitable to explaining/describing differences among groups

[*] Corresponding author: Tel.: +55 81 2126-8430 (x4074), fax: + 55 81 2126-8438
Email addresses: tdpl@cin.ufpe.br (Tarcísio Lucas), tcpbs@cin.ufpe.br (Túlio C. P. B. Silva), rv2@cin.ufpe.br (Renato Vimieiro), tbl@cin.ufpe.br (Teresa B. Ludermir)