A Multi-Objective Memetic Algorithm for the Linguistic Summarization of Time Series [*]

Rita Castillo-Ortega
University of Granada
Granada, Spain
rita@decsai.ugr.es

Nicolás Marín
University of Granada
Granada, Spain
nicm@decsai.ugr.es

Daniel Sánchez
European Centre for Soft Computing
Mieres, Asturias, Spain
daniel.sanchezf@softcomputing.es

Andrea G.B. Tettamanzi
Università degli Studi di Milano
Milan, Italy
andrea.tettamanzi@unimi.it

ABSTRACT

Time series in time domains with a hierarchical structure may be summarized by means of sets of quantified fuzzy sentences of the form "Q of D is A", where Q is a quantifier, D is a linguistic time interval, and A is a linguistic value. Finding concise and accurate summaries that cover the whole time domain is a hard optimization problem, that we solve by proposing a multi-objective memetic algorithm based on NSGA-II with the addition of a number of intelligent mutation operators that apply heuristics to improve solutions.

Track

Genetics-Based Machine Learning

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning—*concept learning*; H.2.8 [**Database Management**]: Database Applications—*data mining*; I.5.1 [**Pattern Recognition**]: Models—*fuzzy set*

General Terms

Algorithms

Keywords

Linguistic Summarization, Multi-Objective Evolutionary Algorithms, Time Series, Multidimensional Data Model, Fuzzy Logic

1. INTRODUCTION

In general, the most common way of expressing time series data is by means of graphical representations, in the form of charts. While very useful and intuitive in many applications, this method may not be very suitable in a number of situations. For instance, if the final users do not have access to machines with adequate graphical capabilities, or the chart is difficult to interpret due to the amount of data or the excessive length of the time span. Another important factor is the granularity employed to show a chart, which sometimes can lead to erroneous interpretations. And, of course, we cannot forget that not all people are able to see properly a graphical representation. Results given in natural language lack all of these problems and, moreover, they can be read out by a voice synthesizer or can be used in a (text-based) information retrieval system. Techniques providing linguistic descriptions of time series are called *time series summarization techniques* in the literature.

This work is based on [1] where the authors use two alternative greedy algorithms to obtain linguistic summaries of rough time-series data fulfilling optimization criteria such as *understandability, brevity, accuracy and coverage*. The greedy technique, based on an incremental construction of the summaries involving locally optimal choices at each step, fails to achieve the full potential of their approach.

The construction of complete, concise and accurate linguistic summaries of time series based on quantified sentences is a hard optimization problem and looks like a perfect fit for a multi-objective evolutionary approach, given the presence of several conflicting criteria and the non-trivial interactions of the various choices that have to be made to obtain a summary (choice of the appropriate quantifiers and of the granularity of the predicates for each sentence, choice of the sentences).

2. LINGUISTIC TIME SUMMARIZATION

The use of powerful tools as fuzzy logic or quantified sentences is well extended through several and different fields, being linguistic summarization one of those ones. The most usual quantified sentences considered in the literature are of the form "Q of X are A" or "Q of D are A", where Q is a linguistic quantifier, X is a (finite) crisp set, and A, D are fuzzy subsets of X. These sentences are called type I and

[*]The research reported in this paper was supported by the Andalusian Government (Junta de Andalucía, Consejería de Innovación, Ciencia y Empresa) under project P07-TIC-03175 *Representación y Manipulación de Objetos Imperfectos en Problemas de Integración de Datos: Una Aplicación a los Almacenes de Objetos de Aprendizaje* and also by the Spanish Government (Science and Innovation Department) under project TIN2009-08296.

GECCO'11, July 12–16, 2011, Dublin, Ireland.
ACM 978-1-4503-0690-4/11/07.

type II sentences, respectively. Using type II sentences our approach will be able to produce sentences like "Most days of the hot season, patient inflow was low or very low".

Our final objective is to obtain a collection of type II quantified sentences. The requirements for this collection of quantified sentences, according to the intuitive idea of summary, are the following: the accomplishment degree of every sentence must be greater than or equal to a user-given threshold τ (accuracy), the set of quantified sentences must be as small as possible (brevity), and the union of the supports of all the time periods in the sentences of the summary must be the whole period (coverage).

Apart from the linguistic partitions D and A and the threshold τ, the user have to select a coherent family of quantifiers, and set Qbound and Gbound (see 3).

3. A MULTI-OBJECTIVE MEMETIC ALG.

NSGA-II [2] works by sorting a population of candidate solution into Pareto fronts, so that non-dominated solutions are in the first front, and applies a niching technique and elitism to improve the population along the entire Pareto front. We have adopted this algorithm and have adapted it to handle some specificities of linguistic summarization.

Our linguistic summary is **represented** by means of a variable-size chromosome, logically divided in genes that encode a single type II quantified sentences. The **initial population** is seeded with individuals with a random number, extracted from an exponential distribution, of sentences whose Q, D, and A are randomly extracted from a uniform distribution. The highly variable (HV) portions of the time series are "masked", so to speak, before starting the EA. Therefore, from EA's viewpoint, it is as if the HV portions of the time series did not exist.

The **objectives** correspond with the quality requirements: *Brevity* of a linguistic summary is computed as the number of quantified sentences that make up the summary. Sentences with group of labels count for as many sentences as there are labels in the group. *Accuracy* is computed for an individual by averaging the accuracies of the sentences that compose it. The accuracy of a single sentence is computed based on the GD method [3]. However, the precision of the quantifier Q is also important and taken into account by means of a specific parameter λ. *Coverage* is computed by counting the number of time points that are covered by at least one sentence in the summary. Since NSGA-II works under the assumption that the objectives are to be minimized, whereas accuracy and coverage are to be maximized, the sign of the latter two criteria is changed to obtain the corresponding objectives for NSGA-II. The **constraints** are handled by adding penalty terms to the relevant objective in case of violation and are enforced by the specialized mutation operators: *Inclusion*, the same time period should not be described by more than one sentence in a summary; *threshold*, the accuracy of a summary must be above a threshold; *Q-bound*, the least strict quantifier allowed in a sentence; and *G-bound*, the maximum label group size allowed in a sentence.

Recombination takes two summaries from the parent population and produces two offspring summaries by uniform crossover (probability p_c). Four **mutation** operators are used: one classical mutation (p_m), which randomly perturbs the genotype simulating transcription errors, and three specialized *intelligent* mutation operators (p_{mi}), which perform meaningful manipulations on the sentences that compose a linguistic summary. These latter have been called *cover* (it looks for non-covered time periods and tries to find suitable labels in the temporal hierarchy to cover them), *split* (it looks for a sentence that can be replaced by more than one new sentences using lower-level labels and it splits it accordingly), and *merge* (it looks for sentences describing adjacent time periods that could be replaced by a single sentence using a higher-level label, and merges them accordingly).

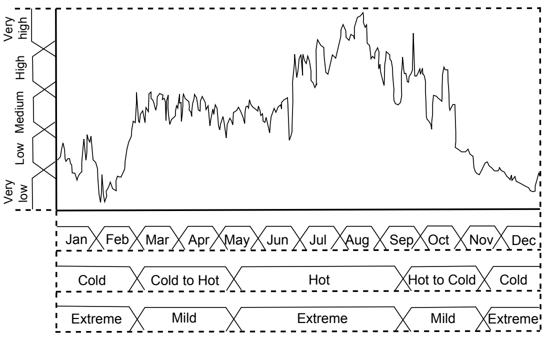

Figure 1: Patient inflow data series.

4. RESULTS

Figure 1 represents the daily patient inflow along a given year to a certain medical centre. The time dimension is hierarchically organized in three fuzzy partitions of the time domain. We have also a fuzzy partition of the inflow basic domain.

The quantifiers are Most of, At least 80%, and At least 70%. $\tau = 0.8$, and $Qbound_i = 3$, $Gbound_i = 2$ in all the levels i of the time dimension. $pop - size = num - generations = 200$, $p_c = 0.5$, $p_m = 0.05$, $p_{mi} = 1$ and $\lambda = 0.7$. As an example within the Pareto's front we have chosen the following solution: *At least 70% of the days with mild weather, the patient inflow is medium or low. Most of the days in September, the patient inflow is high or medium. Most of the days with cold weather, the patient inflow is low or very low. Most of the days in May, the patient inflow is very high or medium. Most of the days in June, the patient inflow is high or medium. Most of the days in July, the patient inflow is high or medium. Most of the days in August, the patient inflow is very high or high.*

Objectives: Brevity: 14, Accuracy: -0.914807 and Coverage: -362. No constraints have been violated.

5. REFERENCES

[1] R. Castillo-Ortega and N. Marín and D. Sánchez, *Linguistic Summary-Based Query Answering on Data Cubes with Time Dimension.* FQAS'09, LNAI, 5822, 560–571, 2009.

[2] K. Deb and S. Agrawal and A. Pratap and T. Meyarivan, *A Fast Elitist Non-Dominated Sorting Genetic Algorithm for Multi-Objective Optimization: NSGA-II.* 849–858, 2000.

[3] M. Delgado and D. Sánchez and M.A. Vila, *Fuzzy Cardinality Based Evaluation of Quantified Sentences.* International Journal of Approximate Reasoning, 23, 23–66, 2000.