Qualitative Object Description: Initial Reports of the Exploration of the Frontier

Igor S. Zwir*
Departamento de Computación
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires, Argentina
zwir@ai.sri.com

Enrique H. Ruspini
Artificial Intelligence Center
SRI International
Menlo Park, California, U.S.A.
ruspini@ai.sri.com

Abstract

The increased ability to access repositories of representations of complex objects, such as biological molecules or financial time series, has not been matched by the availability of tools that permit locating them, visualizing their characteristics, and describe them in terms that are close to the language of the intended users of those data collections. The representation methods and organization schemes employed in the majority of these repositories are based, rather, on considerations related to computational simplicity and efficiency. The observation that humans typically resort to qualitative descriptions of complex objects to describe interesting features of these entities suggests that the automated generation of those descriptions might substantially improve the accessibility and usefulness of repositories of complex objects.

We present progress in a program of research devoted to the automated identification of significant qualitative features of complex objects, the discovery and representation of interesting relations between those features, the generation of structured indexes and textual annotations describing features and their relations, and the discovery of knowledge (or data mining) by analysis of collections of qualitative descriptions.

We focus primarily on methods for the succinct description of interesting features lying in the effective frontier—or the set of all Pareto-optimal solutions—of a multiobjective problem. The multiobjective, generalized clustering, problem being considered is that of extracting features deemed to be interesting from the viewpoint of domain experts. Since, typically, it is easier to derive good explanations of smaller portions of a data object (e.g., a short-lived uptrend) than those of extended portions, any optimization-based formulation of the qualitative feature-identification problem must contend with two conflicting objectives: quality and extent of the description.

We present a Pareto genetic algorithm for the solution of the multiobjective optimization problem. The formulation of this problem is noteworthy because of its treatment of clustering as the isolation of individual clusters (i.e., as opposed to the determination of optimal partitions), its lack of assumptions about the number of clusters, and its ability to deal simultaneously with multiple types of interesting structures. We introduce also an approach based on elimination, simplification, and clustering techniques to summarize and organize the Pareto-optimal solutions lying along the effective frontier. We present experimental results of the application of our methodology to the qualitative description of financial time series.

Keywords: Qualitative Description, Generalized Fuzzy Clustering, Data Mining, Automated Database Annotation, Pareto Optimality, Technical Analysis.

*On leave at Departmento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada, Spain. Currently visiting the Artificial Intelligence Center, SRI International, Menlo Park, California

1 Introduction

The rapid development and implementation of data repositories containing representations of complex objects, such as time series or biological molecules, has not been matched by an increased availability of tools and structures permitting the search of those databases in terms that match the needs and experience of their users. In particular—and in spite of the recent renewed interest in knowledge-discovery techniques (or *data mining*)—there is a dearth of data representation approaches intended to facilitate understanding of the represented objects and related systems. Rather, the structures typically provided to organize and index these collections reflect the convenience of database implementers and their tendency to rely on approaches developed to store simpler, more structured, objects.

The representation of complex objects such as biological molecules as arrays of atoms positions and characteristics, for example, promotes representational accuracy and computational efficiency but fails to provide insights on structural features and functional characteristics of the molecules that may facilitate their understanding. Similarly, financial and economic time series are represented by structures that often conceal rather than reveal important features such as trends or special oscillatory patterns.

This paper presents results of an ongoing research project being carried out at the Artificial Intelligence Center of SRI International. The overall objectives of this program are the extraction of qualitative features from complex objects, the discovery of interesting relations between these features (e.g., inclusion, spatial proximity), the structured and textual annotations of databases on the basis of such discoveries, and, ultimately, the mining of those databases by examination of qualitative properties of the represented objects.

In the rest of this paper, we discuss methods for the identification of qualitative descriptions of significant features of complex objects, their summarization, and interrelation. We briefly describe, in Section 2, our approach to feature extraction, which is based on a treatment of that issue as a generalized clustering problem. In Section 3 we sketch our genetic-algorithm (GA) approach to the solution of this problem, devoting most of our comments, however, to our approach to the summarization and organization of the resulting clusters.

2 Problem

We consider the problem of determining interesting features in a complex object and that of determining interesting relations between them. We assume that "interestingness" is formally defined by means of a family of parameterized models $\mathcal{M} = \{M_\alpha\}$ and by a set of relations between them that are provided beforehand by domain experts.

From this perspective, the problem is formulated as a generalized clustering problem in the sense that extracted subsets meet, to some extent, the requirements imposed by the model collection in the same way that elements of a clustering partition satisfy the constraint that their members be as similar as possible. We need to emphasize, however, that our treatment is more general than that of a typical clustering problem emphasizing the sequential isolation of individual clusters [4], rather than determination of a full clustering. Furthermore, we do not assume a priori knowledge of the total number of clusters, require that the set of all clusters be an exhaustive partition of the complete object, or rely on a single characterization of the notion of cluster.

Our approach follows original ideas of Ruspini [7], as later expanded by Bezdek by introduction of various methods centered upon the notion of prototype [2]. Simple formulation of the feature-identification problem as the optimization of a functional $Q(F, M_\alpha)$—measuring the extent by which the parameterized model M_α fits the subset F of the object being described— would simply result, however, in a large collection of features with small extent, as it is easier to explain, with high accuracy, smaller rather than larger sets. To overcome this problem, the generalized clustering problem was reformulated as a multiobjective optimization problem involving not only a degree of matching Q but also a measure S of the extent of the feature being described.

This formulation of the clustering problem as that of trading off quality of match and extent of the feature being described is conceptually close to a

number of approaches to data explanation, notably those based on the notion of minimum-description length [6] and to our previous efforts toward the solution of the qualitative-description problem by penalty-function methods [9]. Our present approach, however, emphasizes direct measures of modeling utility rather than information-theoretic notions of modeling parsimoniousness. More important, our methods do not rely on weighted combinations of the conflicting objectives measured by Q and S, focusing, instead, on the determination of all situations where it is necessary to consider tradeoffs between those objectives. These situations correspond to the set of *nondominated* solutions of the multiobjective optimization problem, that is, the solutions lying along the *Pareto-optimal* or *effective frontier.*

3 Approach

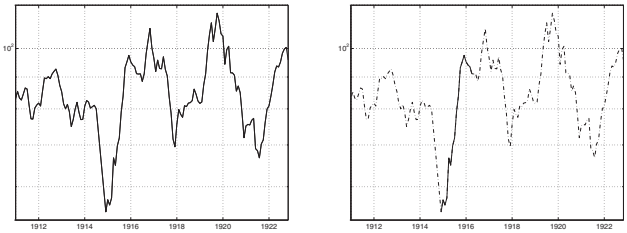
Our approach to the generation of qualitative descriptions of complex objects is based on the generation of the effective frontier of the multiobjective optimization problem and on the subsequent summarization, organization, and description of that set.

3.1 Generating the Effective Frontier

We have developed an algorithm based on the niched Pareto method of Horn, Nafpliotis, and Goldberg [3, 1] to find Pareto-optimal solutions of the multiobjective qualitative-feature identification problem. The output of this algorithm is a set of approximations of nondominated solutions lying in the effective frontier of the optimization problem. Two objectives, corresponding to the quality and extent of a particular description, were considered. Clearly, these objectives are conflicting in the sense that it is easier to explain more accurately smaller than larger subsets of the objects being studied (in our case, corresponding to intervals of a time series). Our implementation of this GA method is described in some detail, together with features of our fuzzy-logic-based approach to model definition, in a recent work [8].

The algorithm was applied to the identification of significant technical-analysis [5] patterns in financial time series. The examples presented in this paper correspond to the identification of three types of such patterns: *uptrend*, *downtrend*, and *heads and shoulders*. Potential solutions of the problem correspond to crisp intervals where the time series matches, to diverse extents, the logic-based definition of those features.

Figure 1 shows some of the qualitative features extracted by our genetic algorithm from a time series, shown in Figure 1(a), of monthly averages of closing values of the Dow-Jones Industrials Average (DJIA) index between 1911 and 1922. Figure 1(b) shows an example of uptrend, Figure 1(c) illustrates an example of downtrend, while Figure 1(d) depicts an example of the head and shoulders pattern.



(a) DJIA (1911–1922) (b) Uptrend

(c) Downtrend (d) Head and Shoulders

Figure 1. Qualitative Feature Extraction from a Time Series

It is important to remark that our formulation—intended to determine features corresponding to complex models—often relies on the logical definition of those models. In our time-series application, for example, uptrend is defined as

$$\texttt{Uptrend} \models (\forall\ \texttt{peaks in interval}\ \texttt{peak} \preceq \texttt{next-peak})\ \texttt{and}\ (\forall\ \texttt{valleys in interval}\ \texttt{valley} \preceq \texttt{next-valley}),$$

where $\preceq$ stands for the fuzzy predicate `approximately smaller or equal`. Simply stated,

in an uptrend interval every peak is (approximately) smaller than or equal to its successor and every valley is (approximately) smaller than or equal its successor. In our application, the ground predicate `approximately smaller or equal` is modeled, using standard conventions, by a trapezoidal function having a soft discontinuity at 0. Application of this formula to a particular interval produces, by application of the combination formulas of fuzzy logic, a number between 0 and 1 describing to what extent the values of the time series in the interval represent an uptrend.

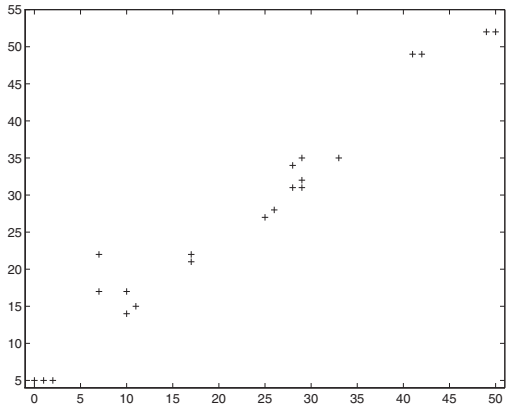
3.2 Describing the Effective Frontier

The output of the Pareto genetic algorithm is a set of approximations of the nondominated solutions of the multiobjective optimization problem. In first analysis it may be thought that, to complete the solution of the qualitative-description problem, the points of this set should be directly related along notions deemed to be interesting by domain experts. In the context of this time-series application, we have assumed that these relations are temporal inclusion (i.e., interval **I** is a subset of interval **I'**), and temporal sequence (i.e., interval **I** temporally follows interval **I'**).

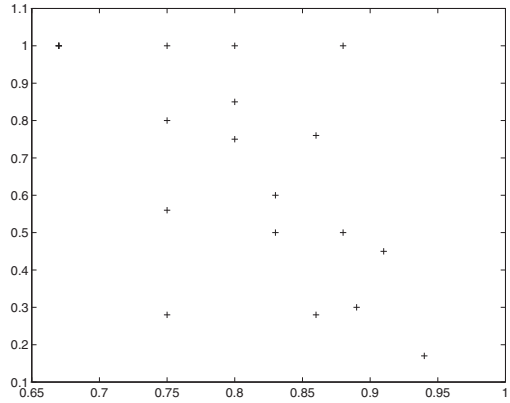
Closer analysis of the resulting solutions, however, indicates that the effective frontier includes a number of solutions that, while different from a strict (or, crisp) viewpoint, might be summarized, as they essentially explain close (or similar) intervals. The set resulting from summarization of the effective frontier provides a more understandable and compact representation of the salient features of the time series.

The nature of the similarity between solutions permitting such summarization is better understood by examination of Figure 2(a), where points in the effective frontier are plotted in the space state (i, I), where i is the leftmost point of the interval and I is its rightmost point. Clearly, the proximity of effective-frontier points in state space suggests clustering of neighboring solutions. The process of effective-frontier description is also simplified by noting, as seen in Figure 2(b), that certain solutions—while lying in the effective frontier and being thus nondominated by their immediate neighbors—are in fact dominated by other, similar, solutions.

Finally, it is also important to note that in (i, I)



(a) Interval Space

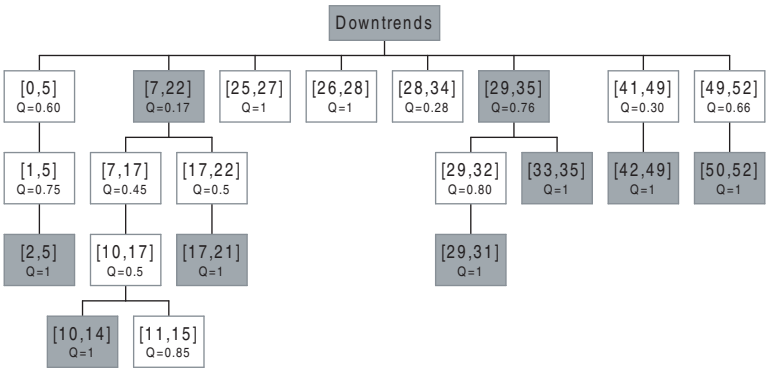


(b) Objective Space (S, Q)

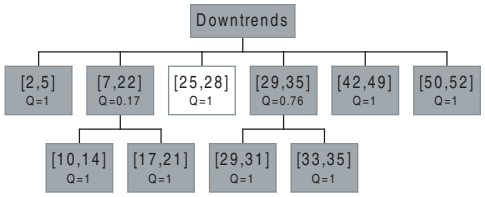
Figure 2. Visualizing the Effective Frontier

space, as shown in Figure 2(a), larger intervals lie close to the upper left-hand corner of the diagram while intervals consisting of a single point lie on the diagonal $i = I$. Solutions having an inclusion are easily visualized in this space as lying on the same perpendicular line to this diagonal (with smaller intervals lying closer to the diagonal).

On the basis of these considerations, we have developed a summarization algorithm intended to provide a compact description of the effective frontier. This algorithm summarizes and relates nondominated solutions by (1) elimination of points being dominated by similar solutions (*fuzzy domination)*, (2) clustering of similar solutions and replacement by prototypes, (3) merging of close, intersecting, intervals with similar values of Q, and (4) hierarchical or-



(a) GA output (Downtrends) organized by inclusion



(b) Summary of the Effective Frontier

Figure 3. Describing the Frontier

ganization of remaining features by inclusion.

This process of summarization and inclusion is illustrated, for downtrend patterns in the DJIA time series, in Figure 3. Figure 3(a) shows an initial organization of all solutions in the effective frontier. Nonshaded boxes indicate solutions to be eliminated (because of poor quality) or summarized (because of similarity to others). The result of these processes and that of merging close intersecting intervals (indicated also by a nonshaded box) can be seen in Figure 3(b).

Because of space constraints, the intervals shown in Figure 3 are represented using an application-dependent interval numbering system (e.g., 7-22) rather than by the true temporal limits of the epoch. Figure 4 illustrates, using actual patterns, the results of the organization and summarization of all interesting epochs for all models (i.e., uptrends, downtrends, and head and shoulders). In addition to showing the conflicting relation between the objectives Q and S, this figure highlights the advantage of consideration of fuzzy matching criteria while displaying also the effectiveness of our summarization techniques.

4 Conclusions

Application of prototype-based, generalized clustering, techniques in combination with genetic algorithm methods to the solution of multiobjective problems provides an effective approach to the identification of interesting features in complex objects. Clustering techniques may also be employed to summarize and produce a compact description of salient features and their relations.

References

[1] T. Bäck, D. Fogel, and Z. Michalewicz, editors. *Handbook of Evolutionary Computation*. Institute of Physics Publishing and Oxford University Press, 1997.

[2] J. C. Bezdek. Fuzzy clustering. In E. H. Ruspini, P. P. Bonissone, and W. Pedrycz, editors, *Handbook of Fuzzy Computation*, chapter F6.2. Institute of Physics Press, 1998.

[3] J. Horn, N. Nafpliotis, and D. Goldberg. A niched Pareto genetic algorithm for multiobjective optimization. In *Proc. First IEEE Conf. on Evolutionary Computation*, pages 82–87, 1994.

[4] R. Krishnapuram and J. Keller. A possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems*, pages 98–110, 1993.

[5] M. J. Pring. *Technical Analysis Explained : The Successful Investor's Guide to Spotting Investment Trends and Turning Points*. McGraw-Hill, 5th edition, 1991.

[6] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific, 1989.

[7] E. H. Ruspini. A new approach to clustering. *Information and Control*, 15(1):22–32, July 1969.

[8] E. H. Ruspini and I. S. Zwir. Automated qualitative description of measurements. In *Proc. 16th IEEE Instrumentation and Measurement Technology Conf.*, 1999.

[9] K. Thranberend and E. H. Ruspini. Subtractive optimization methods for hierarchical fuzzy clustering. In *Proc. 1997 Conf. International Fuzzy Systems Association*, 1997.

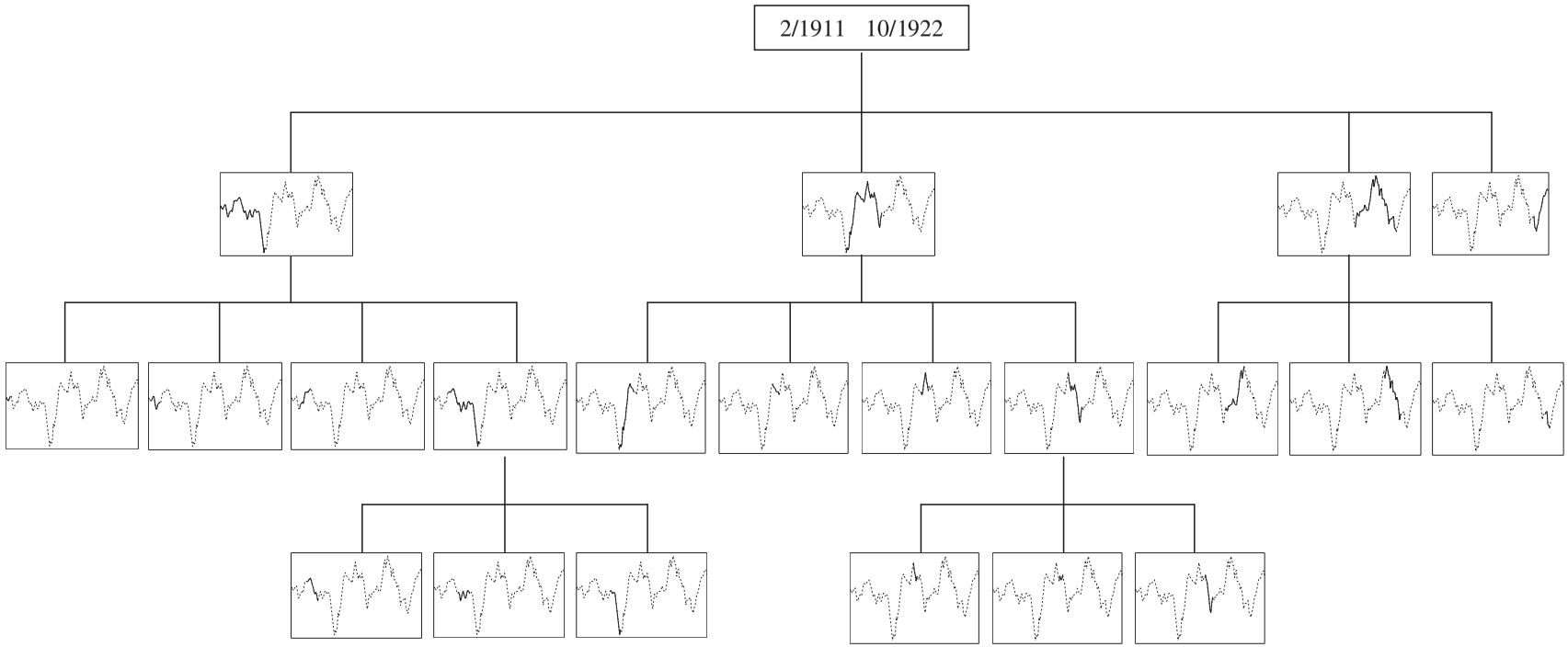


Figure 4. Summarizing the DJIA (1911–1922)