Automated Qualitative Description of Measurements

Enrique H. Ruspini
Artificial Intelligence Center
SRI International
Menlo Park, California 94025, U.S.A.
ruspini@ai.sri.com

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

Abstract

Measurements are usually thought of as precise numbers resulting from observation of a real-world system by means of special sensors or measuring devices. Humans frequently resort, however, to qualitative descriptions—based on the extent by which measurements agree with various models of relevance to the system being observed—to explain the nature and importance of a particular measurement or that of a set of related observations. When describing economic time series, for example, it is customary to point to major qualitative features of the series (e.g., uptrends, downtrends, oscillation patterns) rather than to sequences of precise values. In another important field of study — molecular biology—the positions of atoms in a large, complex molecule may be utilized to produce qualitative descriptions (e.g., shapes of significant molecular structures) that help in the understanding of molecular properties and function.

We present results of ongoing research on methods for the automatic derivation of qualitative descriptions of complex objects. The ultimate goals of these investigations are the development of a methodology for the qualitative representation of complex objects, the systematic search and retrieval of measurements and objects based on those representations, and the discovery of knowledge based on the study of collections of such qualitative descriptions.

Our techniques combine fuzzy logic and evolutionary computation methods to solve optimization problems associated with qualitative description. These methods are noteworthy in that they do not assume prior knowledge of the number of interesting structures, or their extension nor do they require an exhaustive explanation of the object being described. We present results of the application of these methods to the description of financial time series.

*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

We measure real-world systems for two main purposes. The first, and more familiar, of the objectives underlying a measurement action is the identification of the state of a system of known characteristics and structure. In other instances, however, the goal of a measurement activity is the identification of the characteristics of the system, that is, its differentiation among a possible number of structural alternatives. Beyond simply inferring the values of structural parameters, system identification may be broadly described as the specification of the distinguishing characteristics of the system and its behavior (e.g., linear, stable, reactive).

The continued development of large, sophisticated, repositories of knowledge and information has facilitated the accessibility to vast amounts of data—including measurement data—about complex systems and their behavior. The usefulness of these databases is, however, limited by the inability to understand the system-related characteristics of that data (i.e., the knowledge or information conveyed by the data) and by the incapacity to search and retrieve data from these collections on the basis of structural characteristics that are meaningful to humans. In many advanced information repositories, complex objects are modeled by means of structures that promote representational accuracy and computational efficiency but that do not facilitate their analysis and understanding. Similarly, search tools and their supporting structures (e.g., indexes) often fail to provide capabilities to search database contents according to criteria that closely match the experience and need of their users.

A prominent example of this type of objects is the biological molecule, typically represented as a large array of atoms characteristics and positions that do not readily permit the visualization of important characteristics such as surface features or structural patterns. Another example of this class of computational object is the time series where, again, customary descriptions conceal rather than reveal important features such as trends or special temporal patterns.

Substantial gains in the accessibility and overall useful-

ness of these data collections might be achieved by development of automated techniques for the description of objects in terms that are of interest to end users and by the related indexing of digital libraries along lines suggested by these representations.

This paper presents results of research on the identification of qualitative structures in data objects associated with complex systems, their constituent subsystems, and their temporal behavior. These investigations are part of a larger program of research that, beyond discovering such features, seeks to describe their relations (i.e., the *uptrend* preceded the *oscillations*), produce indexes and linguistic descriptions based on discovered features and relations, and discover knowledge (or *data mining*) by exploration of the qualitative descriptions.

2 Problem

We present results of the application of advanced computational techniques to the identification of interesting structures in complex data objects. The data objects to be described are usually represented, nowadays, by various data structures of considerable size and complexity. Data values contained in these structures are typically derived from measurements of system characteristics or behavior. On the basis of knowledge—provided by domain experts—about features of special importance or interest, the methods presented in this paper seek to generate alternative descriptions that permit *identifying* the system as being in a class or classes related to those features.

From such a systems-oriented viewpoint, we may characterize the problem addressed in this work as one of *system identification* or *classification*, seeking to discover interesting substructures so as to be able to describe the system on their basis rather than in terms of the original measurements. From a computational perspective, on the other hand, the problem may be characterized as one of fitting certain models to subsystems or substructures of a complex object and, correspondingly, the principal underlying issues are the isolation of substructures of potential interest and the description of substructure by parameterized models drawn from a catalog provided by domain experts.

2.1 Qualitative Description

Our investigations are oriented toward the automated production of *qualitative descriptions* of complex objects or datasets (e.g., biomolecules or time series). The term "qualitative" is meant to indicate that we intend to identify substructures that match approximately—often measured by some numerical measure of *degree of matching*—an instantiated version of an idealized model derived from expert knowledge.

In our preferred approach to qualitative characterization, we rely on application of fuzzy-logic concepts and techniques to define degree of matching by means of logical expressions containing fuzzy predicates. From these expressions, it is straightforward to produced numerical measures of fitness or matching that basically describe the extent to which the substructure matches the logical model.

Other characterizations of the degree of matching rely on extensions to the fuzzy domain of existing methods to measure compliance of datasets with models (e.g., least-square approximation errors). We are also experimenting with neural network techniques to produce computational structures that learn the required measures from a training set of examples.

2.2 Optimization-based Formulation

Our approach is based on formulation of the qualitative-feature identification problem as a multiobjective optimization. Informally speaking, we seek to pair models—extracted from a given family of parameterized models $\mathcal{M} = \{\mathrm{M}_\alpha(p)\}$, representing interesting structures—with data substructures so that the model does the best possible job of representing that substructure. Whenever that degree of matching is sufficiently high and whenever the data substructure identified is deemed to be sufficiently important (usually determined by its size or extent), then the pair (model, structure) is incorporated as part of the qualitative description of the object. In other words, if the best explanation of a substructure by an instantiated model has a good degree of matching, then the explanation is incorporated as part of the overall qualitative description.

The feature-identification problem is, therefore, formulated as a multiobjective-optimization problem where the objectives are related to the degree of matching between model and substructure and to the size of the substructure being explained. These objectives are typically conflicting as good explanations tend to be limited in extent while those that, conversely, are capable of describing large subsets of the dataset do so poorly. In addition, certain constraints are imposed into the nature of acceptable solutions of the optimization problem to assure that they correspond to true descriptions of meaningful substructures rather than to particular artifacts of the dataset being described. A significant example of this type of constraint is a restriction imposed to prevent involuntary neglect of data points not properly explained by the model (i.e., data mining as the term used to be employed in a pejorative manner in statistics to indicate selective choice of data samples so as to prove ones hypothesis).

3 Methods

Our approach, following an original idea of Ruspini [13], emphasizes the formulation of clustering problems (itself a problem of finding interesting structures in data) as continuous-variable optimization problems over the space of fuzzy subsets of the dataset. Generalizations of this idea are the bases for numerous generalized clustering methods [5]. Consideration of the clustering problem as the determination of optimal fuzzy, rather than conventional, partitions has several methodological and conceptual advantages.

At the conceptual level, availability of fuzzy classifications provides a richer framework for the description of the relations between points and clusters [14] while permitting a less-distorted mapping between the metrics (i.e., similarities) in sample and classification spaces. At the methodological level, the underlying classification problems become continuous optimization problems, which are easier to treat than their discrete counterparts.

A landmark contribution to unsupervised classification methodology was made by Bezdek [3] with his introduction of prototype-based methods based on a generalization of certain classification algorithms of Ball and Hall [2]. The basic idea of summarizing a dataset by a number of representative prototypes, that is, by objects lying in the same space as the sample points, was to be later extended in many significant directions by relaxing the concept in a variety of ways—for example, line segments, ellipsoids [4]. Another development of high relevance to our research has been the generalization of these ideas by Krishnapuram and Keller [10]. This generalization provides the conceptual basis for the identification and extraction of individual clusters (i.e., rather than the determination of a complete clustering or a partition of the dataset into a fixed number of clusters).

Our approach owes much to these generalizations of the notion of clustering, incorporating also specific criteria for the development of measures of cluster quality on the basis of combinations of various measures of explanatory ability and extent. One such formulation, based on an aggregation of conflicting objectives, was recently employed by Thranberend and Ruspini [15] to characterize the problem of isolating linear clusters in econometric time series. This formalism is conceptually close to notions of minimum-description length [12] although it relies on specific measures of explanatory extent rather than on information-theoretic notions of modeling parsimoniousness.

In the treatment presented in this paper we have, instead, relied on formulations based on the extraction of clusters having certain desirable relationships (i.e., Pareto-optimality) between the values of the conflicting objectives. Our preferred approaches to the solution of these problems are evolutionary-computation methods [1].

3.1 Optimization Methods

In the past, genetic algorithms (GA) have been primarily applied to single-objective problems. Multiple-objective problems have been treated by introducing weighted linear combinations of penalty functions. In these cases, the final GA solutions have been found to be very sensitive to small changes in the penalty function coefficients and weighting factors.[1]

Our current work has been based in a different approach oriented toward determining all possible tradeoffs between conflicting objectives. These solutions are said to be *nondominated* in that there are no other solutions that are superior in all objectives (i.e., improvement of one objective results in a lower value for another). The set of nondominated solutions lies on a surface known as the *Pareto optimal frontier*.

We have evaluated various multiobjective approaches [7, 8], focusing eventually on the niched Pareto method of Horn, Nafpliotis, and Goldberg [9]. In this method, binary tournaments, known as *Pareto domination tournaments*, are employed to determine the dominance status of two competitors A and B. If one of the competitors is dominated by a member of the sample while the other competitor is not dominated at all, the nondominated individual wins the tournament. If both or neither are dominated, then fitness sharing is used to determine the winner (whichever has the lower niche count). The sample size is used to control Pareto selection pressure in a manner similar to that employed to regulate tournament size in normal (single-objective) tournament selection.

3.2 Models

Although, conceivably, the models defined by features of interest in our financial time-series analysis problem might have been specified using simple mathematical structures (e.g., linear or quadratic expressions), we sought to represent a wider class of models by introduction of logic-based expressions describing required characteristics of a good fit between model and substructure. Each of these logical expressions is based on a combination of elastic (or fuzzy) predicates essentially that measure—in a [0,1]-scale—the extent by which the structure has some property. The overall logical expression defining the model formed by conjunction of individual requirements is the basis for the definition (by application of truth-combination formulas of fuzzy logic) of functions measuring the quality of the approximation (with 1 being a perfect match and 0 corresponding to a very poor fit).

[1] Previous research on this problem was based on this type of approach [15].

For example, in our time series application:

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

where $\preceq$ stands for the fuzzy predicate `approximately lower or equal`. Simply stated, in an uptrend interval every peak is (approximately) lower than or equal to its successor and every valley is (approximately) lower than or equal its successor. In our application, the ground predicate `approximately lower 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 a financial uptrend.

This logic-based approach is noteworthy in several regards. First, the methodology permits a clear description of the requirements that must be met to qualify a structure as being explained by a model. In addition, whenever the results of a feature-identification experiment do not correspond to the intuitive notion that is being modeled, the logical expression may be readily analyzed and corrected. Furthermore, reliance on logical expressions typically depends on a reasonably small number of predicates combined by means of simple functions. Finally, from a purely experimental viewpoint, the resulting model expressions have the complexity required to determine the applicability of the methodology to a wide class of models.

4 Experimental Results

Our methods were applied to the identification of significant technical-analysis [11] patterns in financial time series. The time series that were analyzed correspond to monthly averages of closing prices of various financial commodities and indexes. In the examples presented in this paper, our GA approach was applied to the monthly averages of closing values of the Dow-Jones Industrial Average index (DJIA) between 1912 and 1922 (Figure 1). A typical experiment involved analysis of time series with 100 to 150 values. We present results corresponding to the identification of three basic structures: *uptrends*, *downtrends*, and *head-and-shoulders* (H&S).

Figure 1. DJIA index (1912–1922)

In our current version of the feature-identification algorithm, the niched Pareto algorithm was employed to determine conventional (i.e., crisp) intervals corresponding to downtrends, uptrends, or H&S intervals. Two objectives were considered. The first objective—*quality of fit*— measures the extent to which the time-series values correspond to a financial uptrend, downtrend, or H&S interval. The second objective—*extent*— measures, through a simple linear functional, the length of the interval being explained. Clearly, these objectives are conflicting in the sense that, typically, it is possible to explain better smaller than larger intervals.

The chromosomes of the GA were coded as a pair of numbers representing an interval of time. A population of size 200 was modified by the GA over a total period of 600 generations. Cross-over probabilities were chosen in the [0.7, 0.9] range, while the mutation probability was 0.1. The niche size (i.e., the proportion of the population where the sharing function is applied) varied from 1% to 10% of the maximum value encoded as an interval end in the chromosomes. The niche size allows the distribution of the population over different solutions in the search space (i.e., it prevents all chromosomes from converging to a few solutions). For computational simplicity, niche counts are calculated on the partly filled next-generation population rather than on the current population [9].

In our experiments we employed a tournament size between 4 and 20 to control the selection pressure. In this regard, is important to note that our evaluations have showed that the niched Pareto algorithm is somewhat sensitive to the selection pressure and to the sharing pressure applied. Small values of the tournament size (close to 1–2% of the population size) results in too many dominated individuals (i.e., a very fuzzy front) while higher values (more than 20%) result in premature convergence to a small portion of the front [6]. Finally, a small percentage of random individuals were introduced in each generation to make the GA more sensible to new zones [6] . These individuals are assigned to zones not yet represented by any of the features that we seek to identify.

In our formulation, uptrends were defined by means of logical expressions based on comparison of successive

peaks and valleys in the time series. The definition of uptrend is a soft definition in the sense that, rather than providing a crisp distinction between matching and nonmatching intervals, it associates a numerical degree of matching between the concept of uptrend and any interval. Figures 2(a) through 2(c) show examples of uptrend intervals determined by our optimization-based approach.

Figure 2(a) Uptrend (12/1914–5/1916)

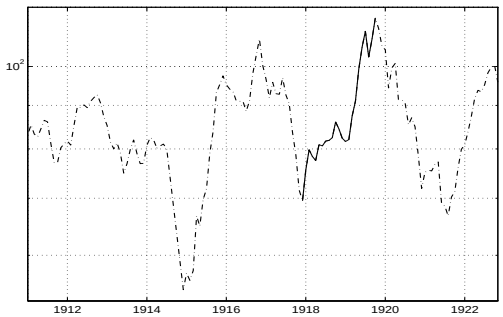

Figure 2(b). Uptrend (12/1917–10/1919)

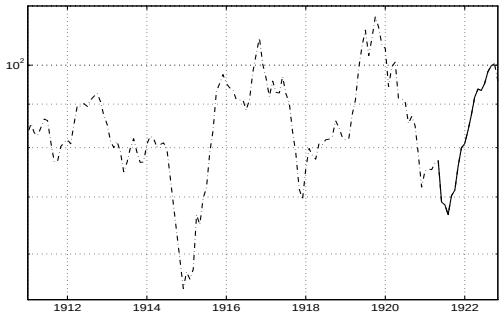

Figure 2(c). Uptrend (5/1921–10/1922)

Downtrends are another example of simple structure that may be defined by means of logical models. As is the case with uptrends, our methods discover (in the Pareto-optimal frontier) downtrend intervals of various extents and quality of fit.

Figures 3(a) through 3(c) show examples of downtrend intervals determined by our optimization-based approach.

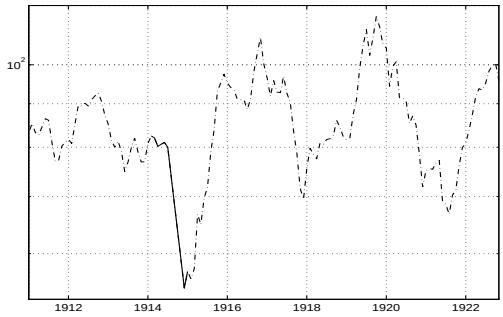

Figure 3(a). Downtrend (2/1914–1/1915)

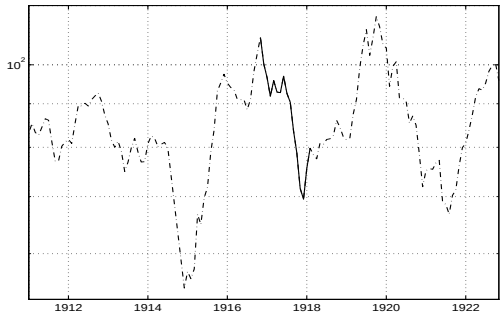

Figure 3(b). Downtrend (11/1916–2/1918)

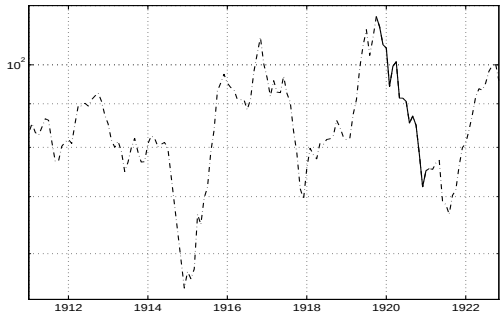

Figure 3(c). Downtrend (10/1919–2/1921)

Finally, we present, in Figures 4(a) through 4(c), examples of a more complex structure—head & shoulders— that were also identified by our approach. Structures such as H&S rely on models that are considerably more complex than those defining (in an approximate fashion) uptrends and downtrends.

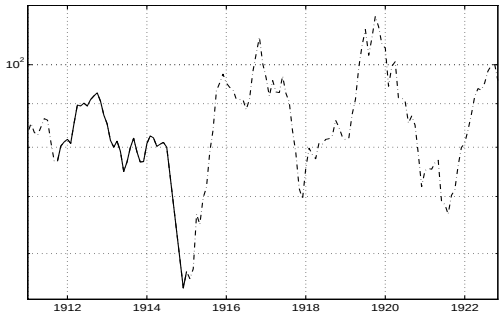

Figure 4(a). H&S (10/1911–1/1915)

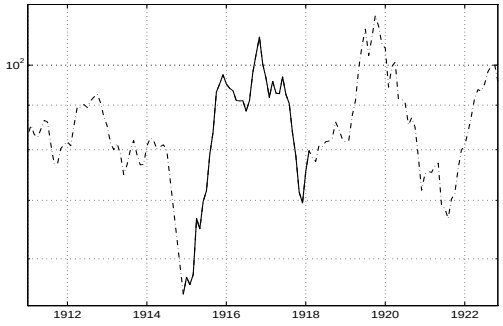

Figure 4(b). H&S (12/1914–2/1918)

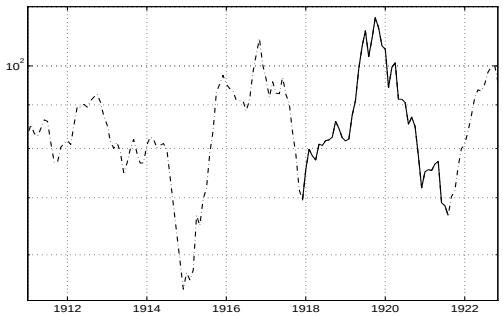

Figure 4(c). H&S (12/1917–8/1921)

5 Conclusions

Generalizations of fundamental ideas that regard cluster analysis as the problem of finding interesting structures in datasets have been applied to describe complex objects and datasets, such as those arising as the result of measurement activities. Application of basic ideas from fuzzy logic permit modeling the problem of extracting qualitative features as a multiobjective optimization problem. The resulting set of extracted features, or *qualitative description*, may be employed to classify and index the corresponding dataset and to identify the underlying system in terms that are meaningful to domain experts.

Experiments in the solution of these multiobjective optimization problems in the context of the description of economical time series by means of genetic-algorithm approaches indicate that the combination of fuzzy logic and evolutionary computation techniques may be successfully applied to the solution of complex qualitative description problems.

We are continuing our research currently seeking methods to summarize and relate the multiple solutions, lying on the effective frontier, produced by our GA-based approach.

References

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

[2] G. Ball and D. Hall. Clustering technique for summarizing multivariate data. *Behav Sci*, 12:153–155, 1967.

[3] J. C. Bezdek. *Fuzzy Mathematics in Pattern Classification*. PhD thesis, Cornell University, 1973.

[4] 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.

[5] J. C. Bezdek and S. Pal, editors. *Fuzzy Models for Pattern Recognition: Methods that Search for Structures in Data*. IEEE Press, 1992.

[6] C. Fonseca and P. Fleming. Multiobjective genetic algorithms made easy: Selection, sharing and mating restriction. In *Proc. the First IEE/IEEE Intl. Conf. on Genetic Algorithms in Engineering Systems*, pages 44–52, 1995.

[7] C. Fonseca and P. J. Fleming. Multiobjective optimization. In T. Bäck, D. Fogel, and Z. Michalewicz, editors, *Handbook of Evolutionary Computation*. Institute of Physics Publishing and Oxford University Press, 1997.

[8] J. Horn. Multicriterion decision making. In T. Bäck, D. Fogel, and Z. Michalewicz, editors, *Handbook of Evolutionary Computation*. Institute of Physics Publishing and Oxford University Press, 1997.

[9] 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.

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

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

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

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

[14] E. H. Ruspini. A theory of fuzzy clustering. In *Proc. 1978 IEEE Intl. Conf. on Decision and Control*. IEEE Press, 1978.

[15] K. Thranberend and E. H. Ruspini. Subtractive optimization methods for hierarchical fuzzy clustering. In *Proc. 1996 Conference International Fuzzy Systems Association*, 1996.