PROPORTIONALITY GRAPHS, UNITS ANALYSIS, AND DOMAIN CONSTRAINTS: IMPROVING THE POWER AND EFFICIENCY OF THE SCIENTIFIC DISCOVERY PROCESS*

Brian Falkenhainer
Department of Computer Science
University of Illinois
1304 W. Springfield Ave, Urbana, IL 61801

ABSTRACT

An important subproblem of scientific discovery is *quantitative discovery,* finding formulas that relate some set (or subset) of a collection of numerical parameters. Current work in quantitative discovery suffers from a lack of efficiency and generality. This paper discusses methods that are efficient and yet general for discovering equations which try to avoid exponential search. Importantly, these methods can derive equations that cover subsets of the data and derive explicit descriptions of when the equations are applicable. These methods are fully implemented in a system named ABACUS which is described and some of its results are presented.

1. INTRODUCTION

A goal of quantitative scientific discovery is to create systems which will be able to propose, more or less on their own, various types of empirical laws which hold for some domain. Such systems should be both general and efficient, should be able to handle irrelevant variables, and should describe, where possible, the limits of applicability for their conclusions. While some progress had been made, most notably the BACON series [Langley 83], existing systems have severe drawbacks. These systems tend to make various assumptions which only enable them to discover laws that hold for roughly all of the given data. They assume that all variables are relevant and also require the user to specify which variables should be treated as independent and which as dependent. This is particularly limiting since in interesting cases we seldom know much about the variables involved.

This paper presents several methods to help overcome these limitations. First, two methods of constraining the search, *proportionality graphs* and *units analysis,* are discussed. Second, a method for deriving several empirical laws that cover different portions of a dataset is introduced that produces a description of when the empirical law is applicable. Finally, some results generated by the program ABACUS are shown.

S. CONSTRAINING 8EARCH

The search for empirical laws can be roughly characterized as searching through the space of equations which can be generated from the variables given in the dataset, stopping whenever a law that accounts for a significant amount of the dataset is found. Two problems confound this search. First, the presence of irrelevant variables in the data can tend to confuse the search process. Second, the process of combining existing terms to form new terms is inherently combinatorial. Attempts have been made to constrain the process using "expectation-driven heuristics" which may use desired formula templates, knowledge about the units of measurement being used, or the domain of the problem in general. An approach which makes no assumptions would be of more value.

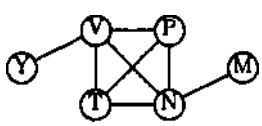
2.1. Proportionality Graphs

Two variables, x and y, may be said to be *qualitatively proportional* if, when other variables are held constant, x rises monotonically as y rises. They are said to be *inversely qualitatively proportional* if x decreases monotonically as y rises. There are four assertions possible as the result of a qualitative proportionality measurement:

prop+(x,y)	- x and y are qualitatively proportional
prop-(x,y)	- x and y arc inversely qualitatively proportional
noprop(x.y)	- x and y are unrelated
prop? (x. y)	- insufficient data to determine if x and y are related

Once all existing variables have been so related, an undirected graph, called the *proportionality graph,* is constructed where nodes represent variables and an edge indicates the existence of a proportionality relation (+ or -) between two variables. It is postulated that, for formulas consisting solely of multiplication and division, the relevant variables in the formula will appear in a cycle in the proportionality graph. This is due to the fact that for formulas of the form (xyz/abc = constant), each variable's value will appear to be dependent upon the other 5. For other formula types, which may or may not adhere to the cycle rule, the most promising area for search would be the largest connected component of the graph.

For example, given data of 6 variables (P, V, N, T, M, Y) where the ideal gas law is the intended relation (PV/NT = 8.32), the proportionality graph might look like:



* This work was supported in part by the National Science Foundation under grant NSF DCR 84-06801 and by the Office of Naval Research under grant no. N00014-82-K-0186.

A simple depth-first search on the relations existing among the set of variables in the cycle {P, V, N, T} produces the desired relation*. No further examination of M or Y are needed for this example. If the quick depth-first search failed, then variables corresponding to edges prop+N.M) and prop±(V,Y) would have been created and search would have continued. This procedure removes irrelevant variables whenever possible and directs the search towards the most promising relations in the data.

2.2. Units Analysis

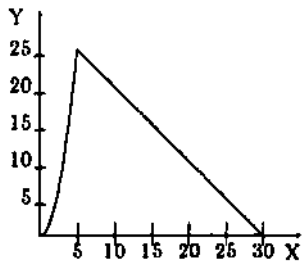
Once it has been decided to create new variables for the relation prop-(x.y), ABACUS tries to form both sum relations (e.g. x+y) and a product relation (xXy) in an effort to create a variable with a constant value. Without further constraint, creating two or more new variables for each proportionality assertion might cause the number of variables in the system to grow exponentially with the search depth. A simple physical constraint drastically reduces the number of combinations that must be considered. For two entities to be added or subtracted, they must be the same type of entity, that is, they must be in the same units. One may divide meters by seconds, but not subtract seconds from meters. Thus, when an attempt is made to form x+y, the units of x and y are examined for compatibility. If x and y have the same units, x+y and x^2+y^2 are created (a need for higher powers has not presented itself as yet). If the units of x are a power of y's units, then y is raised to the appropriate power (e.g. $x + y^2$). Notice we are only testing identity of units and no semantic interpretation is done to guide the search.

3. GENERALIZING THE DISCOVERIES

One of the goals of scientific discovery so far overlooked by past systems is characterizing the domain in which the laws generated are relevant. Only one law could be found for each data set - unlike real life, these programs have not been tested on data sets which include several independent or semi-independent relationships. The ability to classify data according to what mathematical formulae best describes it will be very important as more realistic problems are undertaken.

In the ABACUS system, when the best relation found describes only a portion of the given data, the events covered by this relation are removed from the data and placed in a unique set. The search process begins again to find relationships that describe the remaining events. Entirely new relations may be found this time since the removed variables may have been dominating the relation finding process during the first pass. This process repeats until all of the data has been functionally described or a set of events remains for which no rule can be found. Once all the events are so analyzed, the system attempts to discover what factors present in the data caused set_i to be different from set_j by using an algorithm known as Aq [Michalski 83]. Given event sets such as those created above, Aq will generate a description for each set which covers the examples of that set and none of the examples in any other class. This is called a *discriminant description* in that it may be used to determine to which class a given event belongs. (Aq is described in greater detail in [Michalski 83])

An example will make the role of Aq in the discovery process clearer.



Data corresponding to the above graph was presented to ABACUS. After failing to locate a relation which held for all of the given data, search terminated with $y = x^2$ being the equation which held for the largest percentage of the data . All data covered by this formula was removed from the dataset and search began again. This time the equation describing the line was discovered and found to hold universally for all of the current data. With relations found for the entire dataset, Aq was then called to discriminate between the two sets of events. The results are shown below:

class-a: If [x = 0.10 .. 5.00]
then $x^2 = y$

class-b: If [x = 5.10 .. 30.00]
then x + y = 30

4. IMPLEMENTATION

The concepts presented above are implemented in the ABACUS system, which is written in FranzLisp. ABACUS operates in two stages. First, the given data is examined for empirical rules which hold over different subsets of the events. Once all of the data has been separated into disjoint classes corresponding to each empirical rule, the second phase begins. This stage applies the Aq algorithm to the classes to create discriminate descriptions of the class sets. The class description and empirical rule which describes that class are then combined to form the if-then rules illustrated above.

Since we have already described the role of Aq in some detail, we now describe how the first stage works. First the system searches through the data and produces all possible proportionality assertions. A simple graph traversal routine is then called to return an ordered list of the cycles and connected components of the proportionality graph. A depth-first search is performed on each set in turn, where the

* The correct term here is biconnected component [Aho 76] which refers to all nodes forming a cycle - thus {P V N T} rather than merely {V N T}.

It should be pointed out that the problem of conflicting proportionalities has not been solved here. For the curve, prop*(x,y) holds, while prop-(x,y) holds for the line.. In this example, there were 16 points given for the curve and 7 given for the line. ABACUS is able to solve problems having conflicting proportionalities if one dominates the other, as in this example.

depth of the search for a set is given by the cardinality of the set. Variables, once created, are added to the events list and are never retracted. If a relationship that describes all the data is not found, search proceeds in a best-first manner until such a universal relation is found or a cutoff parameter is reached. If no universal relation is found, the events covered by the relation which summarizes the most events are removed from the data, placed in a unique class set, and all generated variables are thrown out. The process repeats on the remaining data.

4.1. A Few Results

One law which has been discovered by the current ABACUS system is Newton's conservation of linear kinetic energy in a perfectly elastic collision. This law states that the total translational kinetic energy of two bodies remains the same before and after an elastic collision:

$$\tfrac{1}{2}m_1v_1^2 + \tfrac{1}{2}m_2v_2^2 = \tfrac{1}{2}m_1v'^2_1 + \tfrac{1}{2}m_2v'^2_2$$

In this example, ABACUS was presented with data consisting of the eight variables shown above and a nominal collision-type variable which took on values of "elastic" and "inelastic". As can be seen below, the system was able to discover the formula for conservation of kinetic energy. Since this relation only held for about half of the data, further search was attempted on the remaining half, but to no avail. A domain constraint was therefore provided to discriminate the two halves of the data.

class-a: If [collision-type = elastic]
then $m_1*(v_1^2 - v'^2_1) = m_2*(v'^2_2 - v_2^2)$

class-b: If [collision-type = inelastic]
then No formula holds

Another interesting example which helps to show the usefulness of the domain constraint concept involves Coulomb's law of electrical force. The electrical force between two charged objects is given by

$$F = \frac{1}{4\pi\epsilon}\frac{q_1q_2}{r^2}$$

where c is the electrical permittivity of the surrounding medium. ABACUS was presented with data consisting of the force, the two charges, the distance between the objects, and the name of the surrounding medium. The results are given here, where the constant term in each of the Coulomb relations is equal to $4\pi\epsilon$ for that substance.

class-a: If [substance=water]
then $q_2 * q_1 = 8897.352 * F * r^2$

class-b: If [substance=air]
then $q_2 * q_1 = 111.280 * F * r^2$

class-c: If [substance=silicon]
then $q_2 * q_1 = 1312.363 * F * r^2$

class-d: If [substance=germanium]
then $q_2 * q_1 = 1779.015 * F * r^2$

5. CONCLUSION

The ABACUS system is a very useful exercise in simulating a part of the scientific discovery process. It has been designed to be efficient and general and has been shown to be applicable to a variety of chemical and physical laws.

This combination of two completely separate algorithms (the relation finding process and Aq) shows how different approaches to learning can be combined to enable tasks neither could do alone. In experimentation, the pertinent variables are often mixed in with many completely irrelevant variables. With this design, a wide variety of variables may be introduced and decisions of pertinence may be left to the program. With Aq acting as a post-processor, the data can then be examined to answer questions concerning the applicable domain of the results and postulate the reasons for discovered differences in the data.

There are still a number of unresolved problems in quantitative discovery. ABACUS is unable to find laws involving logarithmic or trigonometric relations unless these functions are directly introduced into the data by the user. Also unanswered is the problem of discovering polynomials with coefficients. While a great many laws of physics and chemistry do not contain coefficients in their equations, many important laws do, particularly Newton's laws of motion.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Ryszard Michalski, for his support and advice on this project. I also owe a great deal of gratitude to Ken Forbus, Tony Nowicki, Jeff Becker, and Sue Crane for their time, suggestions, and overall help.

REFERENCES

[1] Aho, A., J.E.Hopcroft, J.D.Ullman, *The Design and Analysis of Computer Algorithms,* Addison-Wesley, 1976.

[2] de Kleer,J., "Qualitative and Quantitative Knowledge in Classical Mechanics," MS Thesis, TR-352, Massachusetts Institute of Technology, Cambridge, MA, 1975.

[3] Falkenhainer,B., "ABACUS: Adding Domain Constraints to Quantitative Scientific Discovery," UIUCDCS-F-84-927, ISO 84-7, University of Illinois, November 1984.

[4] Forbus, K.D., "Qualitative Process Theory," PhD Thesis, TR-798, Massachusetts Institute of Technology, Cambridge, MA, 1984.

[5] Langley,P., G.L.Bradshaw, H.A.Simon, "Rediscovering Chemistry with the Bacon System." In R.S.Michalski, J.G.Carbonell, and T.M.Mitchell (Eds.), *Machine Learning - An Artificial Intelligence Approach.* Tioga Pub., 1983.

[6] Langley,P., G.L.Bradshaw, H.A.Simon, J.Zytkow, "Mechanisms for Qualitative and Quantitative Discovery," *Proceedings of the International Machine Learning Workshop,* Monticello, Illinois, 1983.

[7] Michalski,R.S., "A Comparative Review of Selected Methods for Learning from Examples." In R.S.Michalski, J.G.Carbonell, and T.M.Mitchell (Eds.), *Machine Learning - An Artificial Intelligence Approach.* Tioga Pub., 1983.