Deductive Data Mining: Uncertainty Measures for Banding the Search Space

Dr. Lucian Russell
Expert Decisions Inc. and Argonne National Laboratory
lrussell@xdecision.com

Abstract

One of the interactions between knowledge bases and large databases is the process of augmenting the knowledge base through data mining. Existing techniques for this have serious restrictions, and the problem is compounded by erroneous data. Augmenting existing techniques with deductive data mining could improve the quantity and quality of data mining results. The approach of developing bands of data uncertainty is proposed.

1 Introduction

Computerized knowledge systems require representations of the real world (1); these are models. Having knowledge allows one to search the database for data satisfying "A", and subsequently apply the rule A → B to generate data values for instances of B, or alternatively generate data values F(A) for some function F. In a Platonic ideal world this is enough, but in the real world the ideal representations often are seen only approximately. In Plato's terms we see the ideal as if it were an ephemeral, indistinct projection on the rough wall of a cave. In the case of induction, discovering patterns from data, the models are generated by diligent query of the data. In computerized data, this means querying a database, and the flickers on the wall are corruptions of data. Not paying attention to the flickers can lead to incorrect induction and an erroneous model

2 Knowledge in Imperfect Databases

Although many researchers would pay lip service to the above caution and then proceed to ignore it, the problem cannot be dismissed so easily. There is an underlying assumption among such researchers that there is no problem with data quality, and if there is it is due to faulty operational procedures. It can be fixed, they would argue, by adding data integrity rules to the database. Such assumptions, however, are not validated by reality. First, the CACM (2) devoted a recent issue to the Data Quality problem, so it is real. Second, "business rules", that are supposed to ensure data integrity, change over time, especially in industries subject to governmental regulation. This means that a historical database (e.g. a data warehouse) may contain seemingly inconsistent data.

This problem is a symptom of a more general condition, however. That condition is that any data integrity rules applied to a database are only relevant to the uses made of the database. When an unanticipated use of the data occurs, the data may lack integrity with respect to the new conditions. How many database of American addresses, for example, contain 9 digit zip codes for all entries? The five digit Zip Code is valid for delivery, but mailing discounts require the full 9 digits - a new use.

3 Data Corruption and Knowledge Generation

Data mining seeks to generate knowledge from data, hopefully finding previously unknown relationships. But data mining finds rules that are not true on 100% of the data. In other words some data supports P → Q and some P → ~Q at the same time. This is a difficult problem for knowledge generation, but not one that cannot be overcome. For example, Q may be redefined as a variable Q' with a probability estimated by the relevant frequency for the values Q and Q'.

It would be better, of course, if it could be ascertained whether Q or Q' were a single correct answer. With unclear (e.g. contradictory) information in the database the chance of doing this is not good. It is possible, however, that one of the reasons that relationships are not clearer is that there are errors in the data. This paper examines a method of banding data with certainty measures, and doing data mining first on only the most probable bands. This impacts scalability and the

Proceedings of the 5th KRDB Workshop
Seattle, WA, 31-May-1998
(A. Borgida, V. Chaudhri, M. Staudt, eds.)
http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-10/

1. **Confidence**: %-age of P & Q (i.e. $\mathbf{Q} \wedge \mathbf{P}$) w.r.t. **P** being true.
2. **Goal Coverage:** %-age of P & Q (i.e. $\mathbf{Q} \wedge$ P) w.r.t. $\mathbf{Q} \vee \mathbf{P}$ being true.

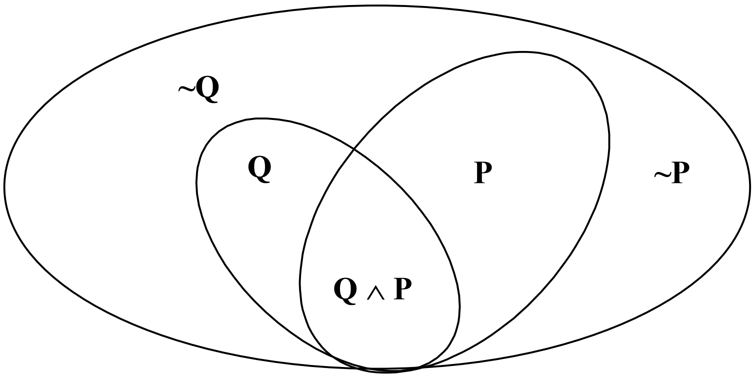

Figure 1: Confidence and Goal Coverage of a Rule

likelihood of finding meaningful relationships among data.[1]

4 The Inductive Paradigm: Rule Discovery

To understand banding, first consider the case of inductive reasoning, specifically the techniques of *rule discovery* (data mining via an unsupervised learning technique). Although it appears to embody the essence of data mining, extracting knowledge from data, it is not all it seems. Let the database have a relational view with attributes (A,B,C … D), and let the goal be to find a logical condition **P** on these attributes and their values that asserts **Q, i.e.** that the database has values (or ranges) for attributes (E,F,G … H). The rule discovered is $\mathbf{P} \rightarrow \mathbf{Q}$ (i.e. **Q** V ~**P** but really only the intersection where $\mathbf{P} \wedge \mathbf{Q}$ is true is of interest). Modeling the logical assertion via sets provides the diagram shown in Figure 1. Because this rule does not hold for all cases, the oval for P is not entirely inside the oval for Q. When performing data mining using rule-discovery techniques, there is a *confidence factor* (**CF**) that quantifies the percentage of the time the rule is true. There is another measure, the *goal coverage* (**GC**) that quantifies the area $\mathbf{P} \wedge \mathbf{Q}$ with respect to the area where ($\mathbf{P} \vee \mathbf{Q}$) is true. Inductive rule generators find large numbers of such rules, especially when thresholds for **CF** and **GC** are small.

It would seem from the above that data mining, perhaps restricted to rules with high confidence, is a good technique for knowledge mining from databases. In practice, however, this pure inductive data mining technique has important non-obvious process dependent decisions.

1. Note that **Q** is not an attribute in the database but rather a logical assertion about database attributes (E,F,G … H) *and their values*.
2. Moreover, due to computational complexity issues it is not about *all* their values, but *a rather small set of subsets of them* (when continuous variables are used, computational considerations require *a relatively small number of ranges of values).*

For purposes of discussion these can be called can be called bands. An example is shown with the three different conclusion values Q_1, Q_2 and Q_3 (e.g. high, medium and low)[2], three bands for the attribute values (E,F,G … H). Rule discovery then looks at what bounds of the variables (A,B,C …D) can be associated with the Q_i. This reduction is what makes rule discovery more computationally tractable, modeling it as a link-analysis or association type of process (see (3) for basic definitions).

In short, inductive rule discovery must be made only on limited, arbitrarily chosen ranges, i.e. *bands*, of data. If alternative ranges are examined the search time increases greatly, and expanding to a full search of the database is computationally intractable. If the wrong values are provided for High Medium and Low there is no provision for trying other values of (E,F,G …H). The knowledge is just not found. In conclusion, knowledge generation in rule discovery systems is performed using bands of values, and induction is made on the data in those bands. We will now look at a technique we call deductive data mining and compare. First, however, we consider how induction and deduction together are used for knowledge generation.

[1] The author recently spent months working with a Fortune 500 company in 1997 where a lack of certainty about the data impeded the investigation.

[2] In the diagram Q1, Q2 and Q3 are NOT meant to indicate nesting but rather partitioning.

5 The KDD Paradigm

According to U. Fayyad in (4) the process called Knowledge Discovery in Databases (KDD) is a discipline that includes "data warehousing, target data selection, cleaning, pre-processing, transformation and reduction, data mining, model selection (or combination), evaluation and interpretation, and finally consolidation and use of the extracted knowledge". Data mining specifically "involves fitting models to or determining patterns from observed data. The fitted models play the role of inferred knowledge" (CACM Vol. 39, No. 11. P 31). Left unsaid is what paradigm is used to bring order to this complex process. However, it does build knowledge from data.

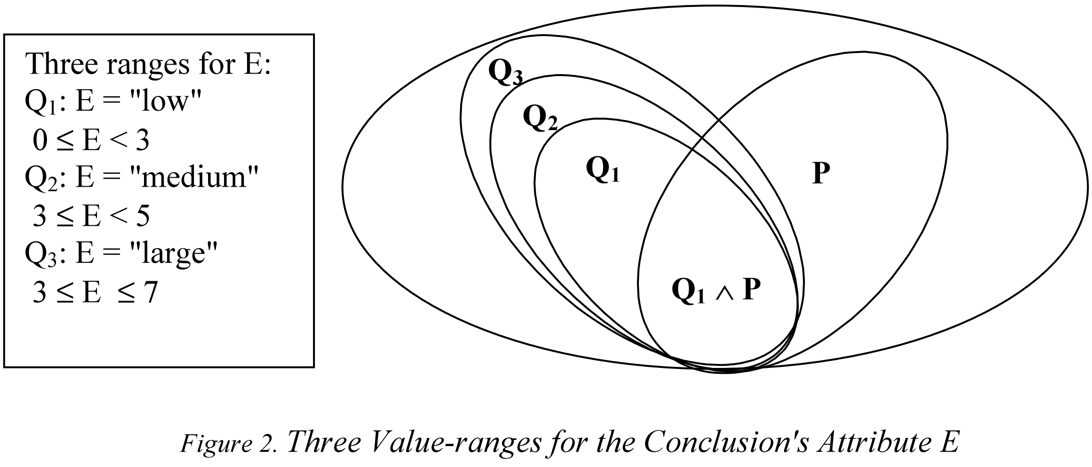

Figure 2. Three Value-ranges for the Conclusion's Attribute E

In actual practice the databases used for data mining have been created as part of the activity shown in Figure 3. There are many hypotheses that are in the mind of the user when the KDD process is begun, and the transformation of variables (e.g. cost/weight instead of cost and weight separately) are initiated because of these hypotheses. Therefore, the long, arduous preparatory work the is performed prior to having a data set for mining, the induction in the KDD process, should be considered as the obverse of an earlier deductive stage of the process. The results from the induction are examined and new working hypotheses formed. One type of deductive transformations, inferring data validity, can be combined with induction to control the region of the search space so as to eliminate spurious results.

6 Deductive Inductive Cycle for Knowledge Generation

To understand deductive data mining, first consider the cycle in Figure 3. In (5) the authors posed an inductive/deductive cycle, and showed how the latter could be supported by the use of the Logical Data Language, LDL++, developed at the Microelectronics and Computer Consortium. This is possible because R. Reiter showed in (6) that in a relational database every query is equivalent to a logical deduction. Thus each query can be considered a proof about a hypothesis, The deductive query finds the data that supports a hypothesis.

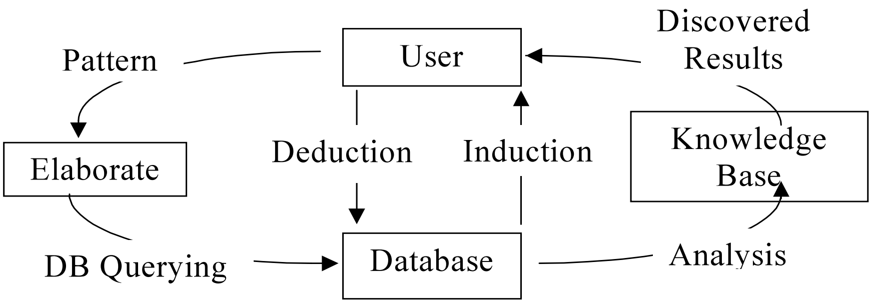

Figure 3: Deductive-inductive cycle for knowledge discovery in databases

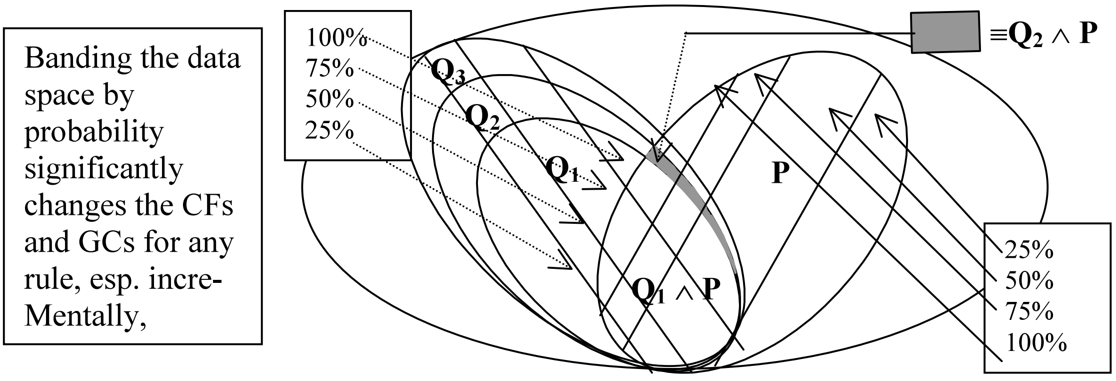
Banding the data space by probability significantly changes the CFs and GCs for any rule, esp. incre-Mentally,
100%
75%
50%
25%
Q3
Q2
Q1
P
Q1 ∧ P
≡Q2 ∧ P
25%
50%
75%
100%

Figure 4: Banding the space by Uncertainty Measures

This technique, however, handles only exact deductive reasoning. Deductive data mining is the technique that extends deductive query techniques to handle uncertain deductive reasoning. This is done by identifying data values that support a chain of deductive reasoning statements, up to a certain level of uncertainty measurement (e.g. probability). This is done by identifying data values that support a chain of deductive reasoning statements, up to a certain level of uncertainty measurement (e.g. probability). Thus each query can be considered a proof about a hypothesis, and each reasoning step involves validating data, at some uncertainty level, needed to support the step. In practice, uncertainty measures of data will be set up in bands (e.g. 0.0-0.5, 0.5-0.7, etc.) which then correspond to database subsets. This searching for data to support hypotheses is also known as retroduction. This directly manages the problem of corrupt data as well. Deductive Data Mining

7 Deductive Data Mining

Consider the variation of Figure 2 shown in Figure 4. An uncertainty measure has been applied to the data in **P** and **Q**, gridding the space of P and Q. The ratios of the **CF** and **GC** are totally altered, depending on the validity thresholds chosen for accepting data as valid. The deductive approach assumes a hypothesis (a database query), and looks at the probability of data supporting it.

The uncertainty information is modeled by adding probability of a value in the database as a column of derived data. Returning to the **P** →**Q** example, doing so subsets the **P** data by row selection, and adds an uncertainty valuation to the values for **Q**, e.g. ((E = 3 or F = 4) ∧ (p(E=3) >0.5) ∧ p(F=4) >0.5))[3].

Notice, however, that expanding the range of the probability search may only slightly expand the search space. This means that as more likely data indicates a search, the scope of the search may increase in a controlled manner, critical to making the approach scalable. This technique is called deductive data mining because it uncovers the data that supports a chain of reasoning. The process is structurally similar because, as with the inductive data mining, the process requires user input. Further, the user input is used to group data values, to provide bands of data.

8 A Military Example

This approach provides the structure in which better knowledge discovery can occur. The data that is most probable can suggest that additional data, at threshold levels below what was originally considered, may have values of interest. From a functional perspective, the flexibility of the adaptive technique is illustrated by Figure 5. Relation E_1 contains radar data from a hilltop and Relation E_2 contains optical image data observed from an airborne platform. The initial **X** is at time T_1 and the final **X** at time T_2. The objects are slow moving (e.g. enemy artillery units). The optical data is subject to distortion, shown as bands of resolution. The radar has

[3] I.e. instead of just looking at simple association rules like $E=c_1 \rightarrow F=c_2$ one should try to deduce rules like $E=c_1 \rightarrow F=c_2$ (probability = c_3) and to do so, one searches a new, slightly richer space ($E=c_1$,0-25%), ($E=c_1$,26-50%),... etc.

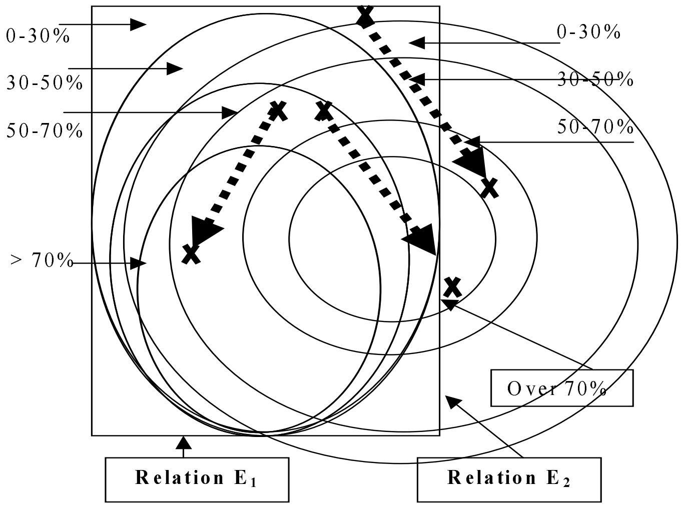

Figure 5: Two types of data need mining

similar bands. Thresholds are initially chosen by the commander, for example 70% for radar and for optical recognition. The data is used for *the hypothesis that there is no threat because the right flank is not threatened.* The percentages are the calibrated likelihood of distinguishing an object correctly. As the data is updated (dotted lines), the leftmost object passes through the different zones, and causes a new hypothesis to be supported: *a threat to the right flank may now exist*. Determining if the potential is an actual threat depends on measuring the optical and radar data relative to their calibration profiles, different scales for different physical realities (Electromagnetic radiation vs. Optics), banding it and mining it.
Consider middle object **X**, which the commander notices at time T_2. The system now initiates dynamic data mining. The system decides to identify the origin of the object, crossing to the database's band for the optical system's mid-level resolution data [50%-70%]. Here a second object **X** on the far right is now identified. Tracking both objects back in time leads to zones of much lower probability for radar and optical resolution. However, the process detects that two enemy units have been moving to the right flank, constituting a threat.

9 Conclusion

Generating Knowledge Bases from Databases through data mining is a technique being applied with much enthusiasm, but with disappointing results. The potential search space is large, and current rule discovery techniques are weak. Current knowledge generation technology, rule generation through inductive data mining, cannot discover knowledge except within bands pre-specified by the user. The technique also is prone to discovering "false positives" of knowledge due to the use of data that is corrupt: i.e. not validated by appropriate data integrity rules. In this context the deductive data mining approach is the equivalent of making lemonade from lemons, in that the data is first conditioned with respect to a query, a working hypothesis. This provides bands of uncertainty in which inductive discoveries may be made more readily. For data mining this means that rather than trying to extract a pattern in a "sea of noise", queries can be used to divide a database into bands, quiet "coves of increasingly probable data". Within the bands of probability the volume of data to be accessed is reduced, so inductive mining techniques have a better chance of uncovering knowledge.

10 Acknowledgement

This work is being supported at Expert Decisions Inc. by Small Business Innovative Research grants from the U.S. Army and U.S. Air Force. The author is on leave from Argonne National Laboratory.

11 Bibliography

1 Mark Stefik,*Introduction to Knowledge Systems*, Morgan Kaufmann, San Mateo, California, 1995
2 Communications of the ACM, February 1998, Volume 41, Number 2
3 Fayaad, U., Piatetsky-Shapiro, G. and Smyth, P. Eds. *Data Mining to Knowledge Discovery: an Overview* in Advances in Knowledge Discovery and Data Mining , MIT Press 1996.
4 Kero, R., Russell L, S. Tsur, W-M Shin, *An Overview of Database Mining,* Proceedings of the KDOOD Workshop, Singapore December 1995.
5 Reiter, R. *A Logical Relational Database Theory*, in On Conceptual Modeling, Brodi M., Mylopoulos J. and Schmidt F.W., Springer Verlag, New York, 1984, pp. 191-238.