Knowledge Discovery from Low Quality Meteorological Databases

C. M. Howard and V. J. Rayward-Smith
School of Information Systems
University of East Anglia, Norwich, NR4 7TJ, UK
{*cmh,vjrs*}*@sys.uea.ac.uk*

April 9, 1998

Abstract

We consider a meteorological application for KDD. The formatting of meteorological problems can yield extremely wide databases, abundant with missing values and unreliable data. We show how feature selection can be applied to remove irrelevant fields from the database thus creating a problem of workable proportions for later stages. Simulated annealing is used to extract rules describing the various outcomes and finally the results are analysed in the context of the problem domain.

1 The Meteorological Domain

Meteorological societies and universities collect vast amounts of data frequently from satellites and weather stations all over the world. Given a collection of datasets, we were asked to examine a sample of such data and look for patterns which may exist between certain geographical locations over time. Similar work has been carried out for some time using standard statistical techniques [2] and occasionally neural networks [18]; part of the aim of the work was to determine whether our approach to data mining could be used to accomplish the same task. The datasets used throughout the work include sea and land surface temperatures, sea level pressures, geomagnetic data and global indicators such as El Niño; recently solar activity has also been suggested as having strong links to long term forecasting [3]. Because of the methods used to collect these datasets during the first half of the 20th Century, a large number of values are missing in approximately half of the database. We find that even where data are available in the the earlier years, there is a degree of unreliability that accompanies it [1]. If a degree of reliability for different periods can be provided by the domain experts, this uncertainty [10, 17] can be built into the data mining algorithm using fuzzy or rough sets [25].

In comparison to some modern databases which require Terabytes of storage, the database described in this paper may seem relatively small. However, it is the shape of the meteorological database that differs most from other databases found in data mining activities. A typical commercial database may have millions of records with a comparatively small number of fields. In contrast the meteorological database begins with over 14,000 fields but with less than 100 records; one record for each year this century. This transposed shape of the data causes problems for many techniques because of the resource requirements and the complexity.

2 The Pre-processing Stage

A number of vital areas that should be addressed in the pre-processing stage are considered below. These areas focus on the problems of data quality, volume and format and suggest methods which can be applied to overcome these.

2.1 Visualisation

Through the entirety of the KDD process it is important to have sufficient understanding of the data. One of the simplest methods of understanding the data appears to be visualisation; a number of packages exist for visualisation in a data mining environment e.g. [22, 24]. The plotting of simple two and three-dimensional graphs can occasionally indicate patterns between fields with very little extra work; although most of the time it is not that simple. The graphing of fields may show erroneous values that exist, possibly due to errors at the time of data input, or it could show that certain *large constant* values may have been used to represent missing values. This visualisation aspect is of particular interest from a meteorology perspective as data (and the location of missing data) can be overlayed with ease onto coastline maps.

2.2 Missing Value Awareness

The database examined in this paper has a significant amount of missing data which not only causes problems when evaluating rules but also to many of the techniques used in the pre-processing stage. Many packages usually require explicit instructions from the user on how to handle missing values and how they are represented. The more naive packages have no facility at all to deal with missing values and those represented by a constant value, usually a large negative number, are taken to be valid data. One of the first places where missing values become a problem is during feature selection, i.e. the identification of important features in the database.

Discarding data The simplest way of dealing with records or fields which contain missing values is to discard them from the database. However, if there are only a small number of fields or a relatively small number of records, discarding those which contain missing values may leave very little of database to work with. This approach may also remove vital information which may be contained within those records or fields; in fact it may be the position of the missing values that is important.

Estimating the missing data A variety of techniques can be used to approximate the values missing from each field [23, 13]. The simplest and quickest method of completing missing values is using the arithmetic mean of the field values although, in meteorological work, climatic averages are favoured which are normally taken over a period of 30 years. Unsupervised clustering methods, such as autoclass–c [15], group similar records according to the input fields; records with missing data can be compared to complete records in the same cluster to estimate missing values. Neural Networks can also be used to complete missing values [6]. Using means to complete the missing values is often unwise, although for the problem at hand, there are too few records in the database for the more involved methods to work sufficiently well.

Deferring the problem One approach to handling missing values is to only use the available values in the field for feature selection. This method may have the disadvantage of making each field a different length but the feature score will be determined only by the correct values in the field and not distorted by placeholders for missing data. When missing values are ignored in this way, the feature score is modified in some way to reflect the use of an incomplete field. The approach of working with missing values rather than completing them so that feature selection can take place, means the problem can be deferred until the data mining stage. In this later phase it will become easier to handle the missing data more accurately since there will be fewer fields as a result of the feature selection. However, not all techniques for feature selection and data mining are available if the missing values are not completed.

Hybrid method The fourth approach to dealing with missing data involves combining the above two methods. The data are temporarily completed, possibly using one of the methods previously described, feature selection is then applied to reduce the number of fields. Once feature selection is complete, the filled-in values are removed in order to return the fields to their original state. The missing values are then handled during the data mining stage.

We recall from [21] that for a database, D, the number of records is defined as $d = |D|$. Then, for any field, f, we define d_f to be the number of records in the database for which that field is defined so $1 \leq d_f \leq d$. If $d_f = d$ the field is said to be complete. The records for which field f is defined form a subset of records, D_f, from the database, D. If f is complete, $D_f = D$. If $D_f = D$ for all fields, the database itself is complete.

2.3 Unreliable Data

In addition to working with databases containing missing values, there is also question about the reliability and quality of some the data. For example, methods for collecting marine temperature data were far more unreliable at the beginning of the 20th Century than now. We might therefore favour rules constructed using records in the latter half of the Century over those constructed using the early half.

If r is a record and f is a field, $f(r)$ will denote the value of field f of r. With uncertain data, each value, v, in the database has an associated confidence level, $c(v)$, where $0 \leq c(v) \leq 1$, and $c(v) = 1$ indicates total confidence in a value and $c(v) = 0$ indicates a missing value. The confidence level is provided by domain experts and could take the form of a step function where, for example in the meteorological context, values up to 1940 may have one level of confidence, values from 1940 to 1960 may have another level and values from 1960 may have total confidence. If we limit our focus to missing values alone, we find we are dealing with the case where $c(v)$ is 0/1 for all v, i.e. a value is either missing or it is available with total confidence. If we are not dealing with the 0/1 case, we can extend the definition of d_f to be

$$d_f = \sum_r c(f(r)).$$

2.4 Discretisation

Discretisation and clustering [8, 14, 16] have many uses in the knowledge discovery process; they can be used to reduce the resource requirements of the data and, more importantly, to simplify the problem. Clustering is the process of grouping together similar real values representing that grouping by a discrete value. Most data mining tools, particularly those used for classification, require the output field to be a discrete value rather than a range of values such as temperature. The Fisher algorithm [5, 7] can be used to discretise the range of values into a specified number of clusters. Using temperature as an example, two clusters may represent high and low, three may represent average with low and high extremes, and so on. Any rules generated in the data mining

stage would therefore describe a range of values represented by a discrete value and the applicability of the rule is controlled by the number of classes in the target field.

2.5 Feature Selection

For most databases with a large number of fields, the feature selection task is possibly the most critical in the data pre-processing stage. The identification of highly predictive fields can prove beneficial in many ways as well as simplifying the problem for the data mining stage. During feature selection, each field is assigned a quality factor determined by a particular algorithm. This score can be used to rank the fields in order of importance and to compare the information provided by one field relative to another.

Feature selection is of particular interest from a meteorological perspective as each field in the database corresponds to a geographical location. By examining the location and time of the high scoring fields, it is possible to construct feature maps. These maps highlight areas of particular interest and show how the relevant fields may shift location with time.

A selection of techniques commonly used for feature selection are described in [4, 9]. For this particular work we have found information gain to be a good indicator of important features. The algorithm is adjusted when missing values exist in the database [20] by scaling the information gain score according to the percentage of data available in the field.

For a field, f, the subset of the database for which the records are defined has been denoted by D_f, clearly $|D_f| = d_f$. The information gain algorithm is applied to this subset and adjusted such that

$$\text{InfoGain}(f) \quad = \quad \frac{d_f}{d} \times \text{InfoGain}(D_f).$$

2.6 Feature Construction

Although in the previous section we try to reduce the number of fields, the complementary action is also advantageous. Constructing a new field from two or more of the original fields can sometimes convey more information than the fields used in isolation.

A particularly useful approach with meteorological data is to calculate differences between fields. The North Atlantic Oscillation index [11] and the El Niño phenomenon which is based on the pressure difference between Darwin and Tahiti [19] have shown to have strong effects on weather conditions worldwide; features such as these could be found using feature construction.

3 The Data Mining Stage

Our approach to the data mining stage is to search for rules of the form $\alpha \Rightarrow \beta$, where α is the precondition of the rule in disjunctive normal form and β is the target postcondition. We then use $\alpha(r)$ to denote that α is true for record r.

In [21] three subsets of the database, D, were defined

$$A = \{\, r \mid \alpha(r) \,\},\ B = \{\, r \mid \beta(r) \,\} \text{ and } C = \{\, r \mid \alpha(r) \wedge \beta(r) \,\} = A \cap B.$$

The values a, b and c were then defined as

$$a = |A|,\ b = |B| \text{ and } c = |C|,$$

and used in the following ratios to measure properties of rules

$$\mathit{Accuracy}(\alpha \Rightarrow \beta) = \frac{c}{a}, \qquad \mathit{Coverage}(\alpha \Rightarrow \beta) = \frac{b}{a}.$$

3.1 Simulated Annealing

Simulated Annealing (SA) is a heuristic search algorithm used for optimisation problems. Previous work [21] has shown SA to be a powerful search tool that can be applied to data mining problems. The SA experiments detailed in this paper use the Templar framework [12]. The framework has been used to produce software to solve data mining problems which generate rules in the required format, using fields with nominal and ordinal values, and allowing for missing values. The evaluation function used to score rules produced in this way is $\lambda c - a$, where λ is a coefficient that controls the accuracy and coverage of the rule.

3.2 SA with Missing and Unreliable Data

When a rule generated by SA is evaluated for a particular record, each inequality in the rule is tested against the corresponding value in the record. If all inequalities hold, the precondition of the rule scores 1 (the corresponding

a is incremented by 1). If the postcondition of the rule also holds when tested on the record, the postcondition scores 1 (the corresponding b is incremented by 1). If both the precondition and the postcondition hold, the entire rule scores 1 (the corresponding c is incremented by 1) otherwise it scores 0. If the value being tested is missing, the inequality cannot be directly evaluated and one of the following choices could be made:

- assume the inequality holds and let the value count towards the precondition score;
- assume the inequality fails which in turn fails the precondition and hence the entire rule;
- adjust the scoring system so that a record is penalised if the corresponding field values in the rule are missing without failing the precondition and entire rule.

Consider a condition, α, defined on k fields. A record, r, is said to totally satisfy that condition if the k fields used in α are all defined in r and the corresponding field values satisfy the condition defined by α.
Consider a predicate α defined on k fields $f_1, ..., f_k$ of the form

$$\alpha \quad = \quad \bigwedge_{i=1}^{k} f_i R_i x_i$$

where $R_i \in \{=, \neq, >, \leq\}$ and $x_i \in \Re$. Defining *true* $\wedge$ *undefined* = *undefined* $\wedge$ *true* = *undefined* and *false* $\wedge$ *undefined* = *undefined* $\wedge$ *false* = *false*, we say a record, r, partially satisfies α iff

$$\bigwedge_{i=1}^{k} f_i(r) R_i x_i \text{ is true.}$$

The strength with which the record, r, satisfies the condition, α, is then defined by

$$s(r, \alpha) \quad = \quad \frac{1}{k} \sum_{i=1}^{k} c(f_i(r)).$$

$s(\alpha, r) = 1$ iff r totally satisfies α, if r does not partially satisfy α, $s(r, \alpha) = 0$. We can now extend the definitions of a, b and c to allow for incomplete and inaccurate data by defining

$$a = \sum_{r \in D} s(r, \alpha), \qquad b = \sum_{r \in D} s(r, \beta), \qquad c = \sum_{r \in D} s(r, \alpha) \cdot s(r, \beta).$$

The definitions of accuracy and coverage can then use the new values of a, b and c. The strategy we are adopting is conservative in the sense that only data occuring in the database can contribute to the value of a, b and c.

4 Results and Analysis

The meteorological database descibed throughout this paper underwent a series of pre-processing and formatting techniques before feature selection took place. The information gain algorithm was applied to the 12,000 fields in the database and the fields ranked in descending order. The best 100 fields were plotted onto coastline diagrams, this highlighted features of geographical areas which were found to have relationships with our selected target area. These findings were passed on to domain experts for validation. These 100 fields were then used in the data mining stage, giving us a problem size which could be processed in a reasonable amount of time.

The 0/1 case for processing missing data has been implemented into our toolkit and number of simulated annealing experiments were carried out which produced rules describing the classes of interest. The rules produced by SA gave a high degree of coverage and accuracy for each of the three target classes; decision tree algorithms were also applied but gave less satisfactory results.

The rule-set has been analysed by experts from the meteorological community and possible justification for the results has been given. It has generally been found that preconditions of the extracted rules have indirect, rather than direct, relevance to the postcondition in terms of climatic features.

When the rules are applied to new data in order to test the results, it is possible that values may now be missing in the test data. The approach used in evaluating the rules during the training stage can also be applied at the testing stage. Effectively, if a particular item is missing when the rule is tested, you can either state there was insufficient data available to test the rule, or evaluate the outcome of the rule and assign it a confidence value based on the proportion of data available.

References

[1] M. Chenoweth. Nineteenth-century marine temperature data: Comments on observing practices and potential biases in marine datasets. *Weather*, 51(8):280–284, 1996.

[2] A. Colman and M. Davey. Linear regression forecast of Central England Temperature for July-August 1996. Technical report, Hadley Centre for Climate Prediction and Research, July 1996.

[3] P. Corbyn. Breakthroughs in long range forecasting (Weather Action Ltd.), 1995.

[4] J. C. W. Debuse and V. J. Rayward-Smith. Feature subset selection within a simulated annealing data mining algorithm. *Journal of Intelligent Information Systems*, 9:57–81, 1997.

[5] J.C.W. Debuse and V.J. Rayward-Smith. One and a half dimensional clustering. In *Proc. of the Conf. on Applied Decision Technology 95*, pages 377–389. UNICOM, 1995.

[6] A. Gupta and M. S. Lam. Estimating missing values using neural networks. *Journal of the Operational Research Society*, 47:229–238, 1996.

[7] J.A. Hartigan. *Clustering Algorithms*. John Wiley & Sons, Inc., first edition, 1975.

[8] K. M. Ho and P. D. Scott. Zeta: A global method for discretisation of continuous variables. In *Proceedings of the KDD '97 Conference*. AAAI, 1997.

[9] C. M. Howard and V. J. Rayward-Smith. Streamlining a meteorological database for knowledge discovery. IEE Digest 97/340, 1997. Colloquium on IT Strategies for Information Overload.

[10] A. Hunter. *Uncertainty in Informations Systems*. McGraw–Hill, 1996.

[11] J. W. Hurrell. Decadal trends in the North Atlantic Oscillation: regional temperatures and precipitation. *Science*, 269:676–679, 1995.

[12] M. S. Jones. The Templar Framework. Technical Report SYS–C9801, University of East Anglia, 1998.

[13] T. L. Walton Jr. Fill-in of missing values in univariate coastal data. *Journal of Applied Statistics*, 23(1):31–39, 1996.

[14] R. Kerber. ChiMerge: Discretization of numeric attributes. In *AAAI-92 Proceedings of Ninth National Conf. on Artificial Intelligence*, pages 123–128, 1992.

[15] K. Lakshminarayan, S. A. Harp, R. Goldman, and T. Samad. Imputation of missing data using machine learning techniques. In U. Fayyad, editor, *Proc. Second Int. Conf. on Knowledge Discovery and Data Mining*. AAAI Press, 1996.

[16] H. Liu and R. Setiono. Chi2: Feature selection and discretization of numeric attributes. In *7th IEEE Int. Conf. on Tools with Artificial Intelligence*, pages 388–391, 1995.

[17] S. McClean and B. Scotney. Role of uncertainty in data mining. In *Proc. of Methods and Tools for Data Mining*. UNICOM, 1997.

[18] T. Miyano and F. Girosi. Forecasting global temperature variations by neural networks. Technical report, Massachusetts Institute of Technology, August 1994.

[19] Britannica Online. El Niño (oceanic phenomenon). http://www.eb.com, March 1998.

[20] J. R. Quinlan. Unknown attribute values in induction.

[21] V.J. Rayward-Smith, J.C.W. Debuse, and B. de la Iglesia. Discovering knowledge in commercial databases using modern heuristic techniques. In *Proceedings of the KDD '96 Conference*. AAAI, 1996.

[22] SGI. MineSet. Silicon Graphics Inc., 1995.

[23] P. K. Sharpe and R. J. Solly. Dealing with missing values in neural network-based diagnostic systems. *Neural Computing and Applications*, 3:73–77, 1995.

[24] G. D. Tattersall and P. R. Limb. Visualisation techniques for data mining. *BT Technology Journal*, 12(4):23–31, 1994.

[25] L. A. Zadeh. From circuit theory to system theory. In *Proc. Institute of Radio Engineers*, volume 50, pages 856–865, 1962.