Automated Discovery of Rules and Exceptions from Distributed Databases Using Aggregates

Rónán Páircéir, Sally McClean and Bryan Scotney

School of Information and Software Engineering, Faculty of Informatics, University of Uster, Cromore Road, Coleraine, BT52 1SA, Northern Ireland.
{r.pairceir, si.mcclean, bw.scotney }@ulst.ac.uk

Abstract. Large amounts of data pose special problems for Knowledge Discovery in Databases. More efficient means are required to ease this problem, and one possibility is the use of sufficient statistics or "aggregates", rather than low level data. This is especially true for Knowledge Discovery from distributed databases. The data of interest is of a similar type to that found in OLAP data cubes and the Data Warehouse. This data is numerical and is described in terms of a number of categorical attributes (Dimensions). Few algorithms to date carry out knowledge discovery on such data. Using aggregate data and accompanying meta-data returned from a number of distributed databases, we use statistical models to identify and highlight relationships between a single numerical attribute and a number of Dimensions. These are initially presented to the user via a graphical interactive middle-ware, which allows drilling down to a more detailed level. On the basis of these relationships, we induce rules in conjunctive normal form. Finally, exceptions to these rules are discovered.

1 Introduction

The evolution of database technology has resulted in the development of efficient tools for manipulating and integrating data. Frequently these data are distributed on different computing systems in various sites. Distributed Database Management Systems provide a superstructure, which integrates either homogeneous or heterogeneous DBMS [1]. In recent years, there has been a convergence between Database Technology and Statistics, partly through the emerging field of Knowledge Discovery in Databases. In Europe this development has been particularly encouraged by the EU Framework IV initiative, with DOSIS projects IDARESA [2] and ADDSIA [3], which retrieve aggregate data from distributed statistical databases via the internet.

In order to alleviate some of the problems associated with mining large sets of low level data, one option is to use a set of sufficient statistics in place of the data itself [4]. In this paper we show how the same results can be obtained by replacing the low level data with our aggregate data. This is especially important in the distributed database situation, where issues associated with slow data transfer and privacy may preclude the transfer of the low level data [5]. The type of data we deal with here is

J.M. Żytkow and J. Rauch (Eds.): PKDD'99, LNAI 1704, pp. 156–164, 1999.

very similar to the multidimensional data stored in the Data Warehouse (DW) [6, 7]. These data consists of two attribute value types: Measures or numerical data, and Dimensions or categorical data. Some of the Dimensions may have an associated hierarchy to specify grouping levels. This paper deals with such data in statistical databases, but should be easily adapted to a distributed DW implementation [8].

In our statistical databases, aggregate data is stored in the form of *Tandem Objects* [9], consisting of two parts: a macro relation and its corresponding meta relations (containing statistical metadata for tasks such as attribute value re-classification and currency conversion). Using this aggregate data, it is possible, with models taken from the field of statistics, to study the relation between a response attribute and one or more explanatory attributes. We use Analysis of Variance (ANOVA) models [10] to discover rules and exceptions from aggregate data retrieved from a number of distributed statistical databases.

Paper Layout

Section 2 contains an extended example. Section 3 shows how the data are retrieved and integrated for final use. The statistical modelling and computation are discussed in section 4, along with the method of displaying the resulting discovered knowledge. Section 5 concludes with a summary and possibilities for further work.

2 An Extended Example

Within our statistical database implementation, the user selects a single Measure and a number of Dimensions from a subject domain for inclusion in the modelling process. The user may restrict the attribute values from any attribute domain, for example, *GENDER= Male*. In this example the Measure selected is *COST* (of Insurance Claim) and the Dimensions of interest are *COUNTRY* {Ireland, England, France}, *REGION* {City, County}, *GENDER* {Male, Female} and *CAR-CLASS* {A, B, C}. A separate distributed database exists for each country.

Once the Measure and Dimensions have been entered, the query is sent to the domain server, where it is decomposed in order to retrieve the aggregate data from the distributed databases. As part of the IDARESA project [2], operators have been developed to create, retrieve and harmonise the aggregate data in the *Tandem* Objects (See Section 3). The Macro relation in the Tandem Object consists of the Dimensions and the single Measure (in this case *COST*), which is summarised within the numerical attributes N, S and SS. S contains the sum of *COST* values aggregated over the Dimension set, SS is the equivalent for sums of squares of *COST* values and N is the count of low level tuples involved in the aggregate. Once the retrieved data have been returned to the domain server and integrated into one Macro relation, the final operation on the data before the statistical analysis is the DATA CUBE operator [11]. Some example tuples from the final Macro relation are shown in Table 2.1.

$$COST_{ijk\ln} = \mu + G_i + P_j + C_k + R(C)_{l(k)} + GP_{ij} + GC_{ik} + GR(C)_{il(k)} + PC_{jk} + PR(C)_{jl(k)} + GPC_{ijk} + \varepsilon_{ijk\ln} \quad (1)$$

Table 2.1. Example tuples from Final Macro relation

COUNTRY	*REGION*	*GENDER*	*CAR-CLASS*	*COST_N*	*COST_S*	*COST_SS*
Ireland	City	Male	A	12000	0.730	43.21
England	County	Female	B	10000	0.517	25.08
All	All	Male	A	72000	4.320	261.23
Ireland	City	All	All	54000	2.850	161.41

The relevant Meta-data retrieved indicates that all the Dimensions are fixed variables for the statistical model, and that a hierarchy exists from REGION → COUNTRY. This information is required to automatically fit the correct ANOVA model. For our illustrative example, the model is shown above in (1).

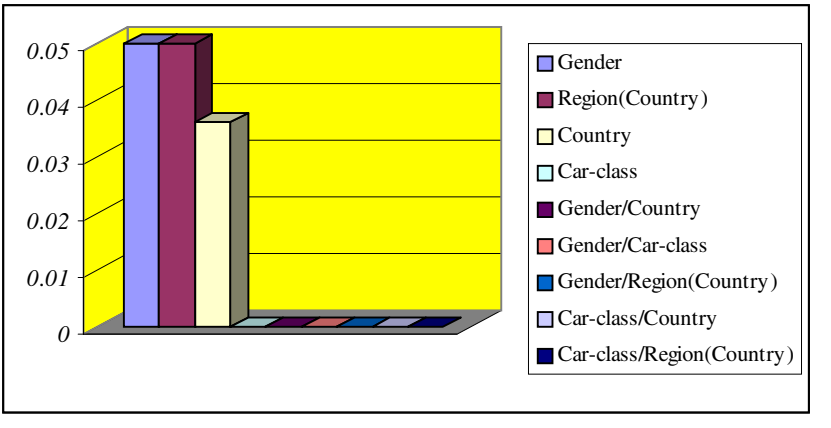

Fig. 2.1. Significant Effects graph for the Insurance example

Once the model parameters have been calculated and validated for appropriateness, the results are presented to the user. The first step involves a graph showing attribute level relationships between the Dimensions and the COST Measure. These relationships (also known as effects) are presented in terms of main Dimension effects, two-, and three- way interaction effects. Only those relationships (effects) that are statistically significant are shown in the graph, with the height of each bar representing the significance of the corresponding effect. The legend contains an entry for all effects, so that the user may drill-down on any one desired. In the Insurance example, *GENDER*, *COUNTRY* and *REGION within COUNTRY* each show a statistically significant relationship with COST, as can be seen from the Significant Effects graph in Figure 2.1. None of the three-way effects (e.g. *GENDER/REGION(COUNTRY)*) have a statistically significant relationship with the COST Measure.

The user can interact with this graphical representation. By clicking on a particular bar or effect in the legend of the graph, the user can view a breakdown of COST values for that effect, either in a table or a graphical format. This illustrates to the

user, at a more detailed level, the relationship between an attribute's domain values and the COST Measure. These are conveyed in terms of deviations from the overall mean, in descending order. In this way, the user guides what details he wants to look at, from a high level attribute view to lower more detailed levels. A graph of the breakdown of attribute values for *GENDER* is shown in Figure 2.2. From this it can be seen that there is a large difference between COST claims for Males and Females.

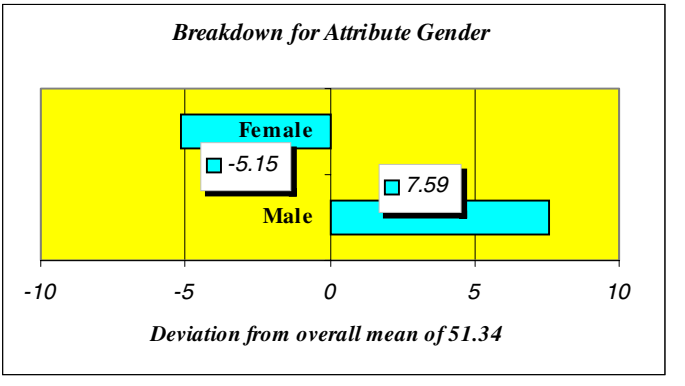

Fig. 2.2. Deviations from mean for GENDER values

On the basis of these relationships, rules in conjunctive normal form (CNF) are constructed. The rules involving GENDER are shown in (2) and (3) below. Based on the records in the databases, we can say statistically at a 95% level of confidence that the true COST lies within the values shown in the rule consequent.

GENDER{Male}	→	COST between {57.63} and {60.23}	(2)
GENDER{Female}	→	COST between {44.63} and {47.75}	(3)

The final step involves presenting to the user any attribute value combinations at aggregate levels which deviate from the high level rules discovered. For example, a group of 9,000 people represented by the following conjunction of attribute values (4) represents an exception to the high level rules:

COUNTRY{Ireland} ∧ GENDER{Female} ∧ REGION{City}			
→	**ACTUAL VALUE**	COST between {50.12} and {57.24}	
→	**EXPECTED VALUE**	COST between {41.00} and {48.12}	(4)

This can be seen to be an exception, as the corresponding Expected and Actual COST ranges do not overlap. The information in this exception rule may be of interest for example in setting prices for Insurance for Females. Before making any decisions, this exception should be investigated in detail. We find such exceptions at aggregate levels only. It is not possible at this stage to study exceptions for low-level values as

these are resident at the different distributed databases, and in many situations privacy issues prevent analysis at this level in any case.

3 Aggregate Data Retrieval and Integration

The data at any one site may consist of low level "micro" data and/or aggregate "macro" data, along with accompanying statistical metadata (required for example for harmonisation of the data at the domain server). This view of micro and macro data is similar to the base data and materialised views held in the Data Warehouse [7]. In addition, textual (passive) metadata for use in documentation are held in an object-oriented database. An integrated relational strategy for micro and macrodata is provided by the MIMAD model [9] which is used in our implementation. To retrieve aggregate data from the distributed data sites, IDARESA has developed a complete set of operators to work with *Tandem Objects* [2]. Within a Tandem Object, a *macro relation* $\mathbf{R}$ $<C_1,\ldots C_n; S_1, \ldots S_m>$ describes a set of macro objects (statistical tables) where $C_1,..C_n$ represent n Dimensions and $S_1,\ldots S_m$ are m summary attributes (N, S and SS) which summarise an underlying Measure. The IDARESA operators are implemented using SQL which operates simultaneously on a Macro relation and on its accompanying meta relations. In this way, whenever a macro relation is altered by an operator, the accompanying meta relations are always adjusted appropriately.

The summary attributes in the macro relation form a set of "sufficient statistics" in the form of count (N), sum (S) and sums of squares (SS) for the desired aggregate function. An important concept is the additive property of these summary attributes [9] defined as follows:

$$\sigma (\alpha \text{ UNION } \beta) = \sigma (\alpha) + \sigma (\beta) \tag{5}$$

where α and β are macro relations which are macro compatible and $\sigma()$ is an application of a summary attribute function (e.g. SUM) over the Measure in α and β. Using the three summary attributes, it is possible to compute a large number of statistical procedures [9], including ANOVA models. Thus it is possible to combine aggregates over these summary statistics at a central site for our knowledge discovery purposes.

The user query is decomposed by a Query Agent which sends out Tandem Object requests to the relevant distributed sites. If the data at a site is in the micro data format an IDARESA operator called MIMAC (Micro to Macro Create) is used to construct a Tandem Object with the required Measure and Dimensions, along with accompanying meta relations. If the data are already in a macro data format, IDARESA operators TAP (Tandem Project) and TASEL (Tandem Select) are used to obtain the required Tandem Object. Once this initial Tandem Object has been created at each site, operators TAREC (Tandem Reclassify) and TACO (Tandem Convert) may be applied to the macro relations using information in the meta relations. TAREC can be used in two ways: the first is in translating attribute domain values to a single common language for all distributed macro relations (e.g. changing French words for male and female in the GENDER attribute to English).; the second use is on reclassi-

fying a Dimension's domain values so that all the macro relations contain attributes with the same domain set (e.g. the French database might classify Employed as "Part-time" and "Full-time" separately. These need to be reclassified and aggregated to the value "Employed" which is the appropriate classification used by the other Countries involved in the query). The operator TACO is used to convert the Measure summary attributes to a common scale for all sites (e.g. converting COST from local currency to ECU for each site using conversion information in the meta relations).

The final harmonised Tandem Object from each site is communicated to the Domain Server. The Macro relations are now *Macro compatible* [2] and can therefore be integrated into a single aggregate macro relation using the TANINT (Tandem Integration) operator. The meta relations are also integrated accordingly. The final task is to apply the DATA CUBE operator [11] to the Macro relation. The data is now in a suitable format for the statistical modelling.

3.1 Implementation Issues

In our prototype the micro data and Tandem Objects are stored in MS SQL Server. Access to remote distributed servers is achieved via the Internet in a Java environment. A well acknowledged three tier architecture has been adopted for the design. The logical structure consists of a front-end user (the client), a back-end user (the server), and middleware which maintains communication between the client and the server. The distributed computing middleware capability called remote method invocation (RMI) is used here. A query is transformed into a series of nested IDARESA operators and passed to the Query Agent for assembly into SQL and execution.

4 Statistical Modelling and Results Display

ANOVA models [10] are versatile statistical tools for studying the relation between a numerical attribute and a number of explanatory attributes. Two factors have enabled us to construct these models from our distributed data. The first is the fact that we can combine distributed primitive summary attributes (N, S and SS) from each distributed database seamlessly using the MIMAD model and IDARESA operators described in Section 3. The second factor is that it is possible to use these attributes to compute the coefficients of an ANOVA model in a computationally efficient way.

The ANOVA model coefficients also enable us to identify exceptions in the aggregate data. The term "exception" here is defined as an aggregate Measure value which differs in a statistically significant manner from its expected value calculated from the model. While it is not the focus of this paper to detail the ANOVA computations, a brief description follows. The simplest example of an ANOVA model is shown in equation (6), similar to the model in equation (1) which contains more Dimensions and a hierarchy between the Dimensions COUNTRY and REGION.

In equation (6), $Measure_{ijk}$ represents a numerical Measure value corresponding to $value_i$ of Dimension A and $value_j$ of Dimension B. k represents the k^{th} example or replicate for this Dimension set. The μ term in the model represents the overall aver-

age or mean value for the Measure. The *A* and *B* single Dimension terms are used in the model to see if these Dimensions have a relationship (Main effect) with the *Measure*. The *(AB)* term, representing a 2-way interaction effect between Dimensions A and B, is used to see if there is a relationship between the Measure and values for Dimension *A*, which hold only when Dimension *B* has a certain value. The final term in the model is an error term, which is used to see if any relationships are real in a statistically significant way.

$$Measure_{ijk} = \mu + A_i + B_j + (AB)_{ij} + \varepsilon_{ijk} \tag{6}$$

In order to discover exceptions at aggregate levels, the expected value for a particular Measure value as calculated from the model, is subtracted from the actual Measure value. If the difference is statistically significant in terms of the model error, this value is deemed to be an exception. When calculating an expected value for $Measure_{ijk}$, the model reduces nicely to the average of the *k* values where A=i and B=j, saving considerable computing time in the calculation of exceptions. The model reduces similarly in calculating an exception at any aggregate level (e.g. the expected Measure value for aggregate GENDER{Male} and COUNTRY {Ireland} is simply the average over all tuples with these attribute values). It is important to note that if an interaction effect (e.g. AB) is deemed to be statistically significant, then the main effects involved in this interaction effect (A and B) are disregarded and all the focus centers on the interaction effect. In such a situation, when effects are converted to CNF rules, main effects based on a significant interaction effect are not shown.

In our ANOVA model implementations, we do not model higher than 3-way interaction effects as these are seldom if ever significant [10].

4.1 Presentation of Results

The first step in the results presentation is at the attribute level based on the statistically significant Main and Interaction Effects. Statistical packages present ANOVA results in a complicated table suitable for statisticians. Our approach summarises the main details of this output in a format more suited to a user not overly familiar with statistical modelling and analysis. We present the statistically significant effects in an interactive graphical way, as shown in Figure 2.1. The scale of the graph is the probability that an effect is real. Only those effects significant above a 95% statistical level are shown. The more significant an effect, the stronger the relationship between Dimensions in the effect and the Measure.

As a drill-down step from the attribute level, the user can interact with the graph to obtain a breakdown of Measure value means for any effect. This allows the user to understand an effect's relationship with the Measure in greater detail. The user can view this breakdown either graphically, as shown in Figure 2.2. or in a table format. The breakdown consists of the mean Measure deviation values from the overall Measure mean, for the corresponding effect's Dimension values (e.g. Figure 2.2 shows that the mean COST for GENDER{Male} deviates from the overall mean of 51.34 by +7.59 units). Showing the breakdown as deviations from the overall mean

facilitates easy comparison of the different Measure means. The significant effects are next converted into a set of rules in conjunctive normal form, with an associated range within which we can statistically state that the true Measure value lies. This range is based on a statistical confidence interval. This set of rules in CNF summarises the knowledge discovered using the ANOVA analysis.

The final pieces of knowledge which are automatically presented to the user, are the exceptions to the discovered rules. These are Measure values corresponding to all the different Dimension sets at the aggregate level, which differ in a statistically significant way from their expected ANOVA model values. An example of an exception is (4) in Section 2. These are also presented in CNF, with their expected and actual range values.

One factor which is also important to a user interested in finding exceptions, is to know in what way they are exceptions. This is possible through an examination of the rules which are relevant to the exception. For the example in Section 2, assume that (2) and (3) are the only significant rules induced. In order to see why (4) is an exception, we look at rules which are related to it. We define a rule and an exception to be related if the rule antecedent is nested within the exception antecedent. In this case the antecedent in rule (3) GENDER{Female} is nested in the exception (4). Comparing the Measure value range for the rule {44.6 - 47.75} with that of the exception {50.12 - 57.24}, it can be seen that they do not overlap. Therefore it can be stated in this simple illustration that GENDER is in some sense a cause of exception (4). This conveys more knowledge to the user about the exception. Further work is required on this last concept to automate the process in some suitable way.

4.2 Related Work

In the area of supervised learning, a lot of research has been carried out on the discovery of rules in CNF, and some work is proceeding on the discovery of exceptions and deviations for this type of data [13, 14]. A lot less work in the knowledge discovery area has been carried out in relation to a numerical attribute described in terms of categorical attributes. Some closely related research involves a paper on exploring exceptions in OLAP data cubes [14]. The authors there use an ANOVA model to enable a user to navigate through exceptions using an OLAP tool, highlighting drill-down options which contain interesting exceptions. Their work bears similarity only to the exception part of our results presentation, whereas we present exceptions to our rules at aggregate levels in CNF. Some work on knowledge discovery in distributed databases has been carried out in [5, 15].

5 Summary and Further Work

Using aggregate data and accompanying meta-data returned from a number of distributed databases, we used ANOVA models to identify and highlight relationships between a single numerical attribute and a number of Dimensions. On the basis of these relationships which are presented to the user in a graphical fashion, rules were induced in conjunctive normal form and exceptions to these rules were discovered.

Further work can be carried out on the application of Aggregate data to other knowledge discovery techniques applied to the distributed setting, conversion of our rules into linguistic summaries of the relationships and exceptions and investigation of models which include a mix of Measures and Dimensions.

References

1. Bell, D., Grimson, J.: Distributed database systems. Wokingham : Addison-Wesley, (1992)
2. M^{c}Clean S., Grossman, W. and Froeschl, K.: Towards Metadata-Guided Distributed Statistical Processing. NTTS'98 Sorrento, Italy (1998): 327-332
3. Lamb, J., Hewer, A., Karali, I., Kurki-Suonio, M., Murtagh, F., Scotney, B., Smart C., Pragash K. : The ADDSIA (Access to Distributed Databases for Statistical Information and Analysis) Project. DOSIS project paper 1, NTTS-98, Sorrento, Italy. 1-20 (1998)
4. Graefe, G, Fayyad, U., Chaudhuri, S.: On the Efficient Gathering of Sufficient Statistics for Classification from Large SQL Databases. KDD (1998): 204-208
5. Aronis, J., Kolluri, V., Provost, F., and Buchanan, B.: The WoRLD: Knowledge Discovery from multiple distributed databases. In Proc FLAIRS'97 (1997)
6. Chaudhuri, S., Dayal, U.: An Overview of Data Warehousing and OLAP Technology. SIGMOD Record 26(1): 65-74 (1997)
7. Shoshani, A.: OLAP and Statistical Databases: Similarities and Differences. PODS 97: 185-196 (1997)
8. Albrecht, J. and Lehrner, W.: On-Line Analytical Processing in Distributed Data Warehouses. International Database Engineering and Applications Symposium (IDEAS'98), Cardiff, Wales, U.K (1998)
9. Sadreddini MH, Bell D., and McClean SI.: A Model for integration of Raw Data and Aggregate Views in Heterogeneous Statistical Databases. Database Technology vol 4,no 2, 115-127 (1991).
10. Neter, J.: Applied linear statistical models. - 3rd ed. - Chicago, Ill.; London: Irwin, (1996).
11. Jim Gray, Adam Bosworth, Andrew Layman, Hamid Pirahesh: Data Cube: A Relational Aggregation Operator Generalizing Group-By,Cross-Tab, and Sub-Total. ICDE 1996: 152-159 (1996)
12. Liu, H., Lu, h., Feng, L. and Hussain, F.: Efficient Search of Reliable Exceptions. PAKDD 99 Beijing, China (1999)
13. Arning,A., Agrawal, R. and Raghavan, P.: A linear Method for Deviation Detection in Large Databases KDD, Portland, Oregon, USA (1996)
14. Sarawagi, S., Agrawal, R., Megiddo, N.: Discovery-Driven Exploration of OLAP Data Cubes. EDBT 98: 168-182 (1998)
15. Ras, Z., Zytkow J.:Discovery of Equations and the Shared Operational Semantics in Distributed Autonomous Databases. PAKDD99 Beijing, China (1999)