Using Hierarchies, Aggregates and Statistical models to discover Knowledge from Distributed Databases

Rónán Páircéir, Sally McClean and Bryan Scotney

School of Information and Software Engineering, Faculty of Informatics,

University of Ulster, Cromore Road, Coleraine, BT52 1SA, Northern Ireland.

{r.pairceir, si.mcclean, bw.scotney}@ulst.ac.uk

Abstract

Data Warehouses and statistical databases (Shoshani 1997) contain both numerical attributes (measures) and categorical attributes (dimensions). These data are often stored within a relational database with an associated hierarchical structure. There are few algorithms to date that explicitly exploit this hierarchical structure when carrying out knowledge discovery on such data. We look at a number of aspects of knowledge discovery from a set of databases distributed over the internet including the following:

- Discovery of statistical relationships, rules and exceptions from hierarchically structured data which may contain heterogeneous and non-independent instances;
- Use of aggregates as a set of sufficient statistics in place of base data for efficient model computation;
- Leveraging the power of a relational database system for efficient computation of sufficient statistics;
- Use of statistical metadata to aid distributed data integration and knowledge discovery.

Introduction

Frequently data which are stored in transaction databases, statistical databases and data warehouses are contained within a hierarchical structure. A typical example is supermarket data where individual products can be grouped within product groupings, within food categories. This type of data has been widely used recently for much of the work in the area of discovering association rule at different levels of hierarchal transaction data (Srikant et al. 1995). However little work has been carried out to date in the area of knowledge discovery from hierarchical data which consists of both numerical (measures) and categorical (dimensions) attributes. This type of data is typically found in statistical databases and data warehouses.

Attempting to carry out knowledge discovery from such data when the data are distributed in a number of databases over the internet adds an extra dimension to the problem. Aggregates are especially appropriate for Knowledge Discovery from distributed databases for reasons of confidentiality and efficiency. Using aggregate data and accompanying metadata retrieved from a set of distributed database, we use statistical models to identify and present relationships between a single numerical attribute and a combination of other attributes at various levels of the hierarchy. On the basis of these relationships and interactions, rules in conjunctive normal form are induced.

We illustrate this work with vehicle insurance data. The task is to discover relationships between *vehicle insurance claim costs* and a host of other attributes at different levels of a hierarchy. The attributes of interest are shown in Table 2 along with their associated level in the hierarchy.

Distributed Data

When carrying out knowledge discovery from a number of databases distributed over the internet, it may be either too expensive or prohibited by confidentiality constraints to communicate base (micro) data across the network (Provost 1998). One strategy in this instance is to use meta-learning (Stolfo et al. 1997), whereby a model is learned at each distributed site and sent in place of the base data to a central site to be used to build a combined model. We overcome this problem by communicating aggregate data to a central site rather than the base data. At the central site, the aggregate data are harmonised, integrated and then used as a set of sufficient statistics to reveal statistical relationships in the data.

In (Páircéir et al. 1999) we presented an infrastructure based on a European DOSIS project ADDSIA (McClean et al 2000), for integrating aggregate data from horizontally partitioned base data held in a number of statistical databases distributed over the internet. In ADDSIA, a MAMED object consists of an aggregate dataset (formed from the base data at a site) with associated active statistical metadata which is held in relational tables. A set of MAMED operators have been developed to create the MAMED objects at each site. Data from the different sites (held in different countries in this instance) may contain attribute domain value mismatches which need to be harmonised before the data can be integrated for data mining. For example, domain values of a dimension attribute may be recorded at different levels of granularity

on different sites, or a measure attribute may record costs in different currencies on the different sites.

Aggregate data harmonisation and integration at the central site are accomplished using further MAMED operators, with the aid of the active statistical metadata in the MAMED objects. The statistical metadata also contain data which are required for statistical modeling, including information about the hierarchical structure of the data. At the central site the associated statistical metadata tables are also integrated and can be viewed by the user.
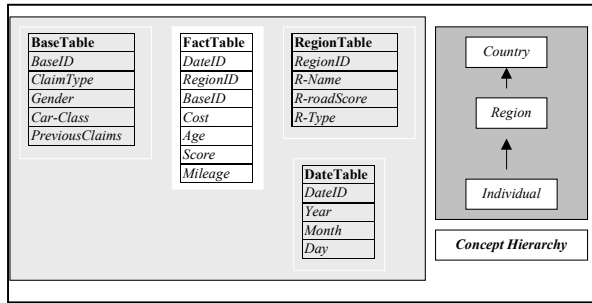

Figure 1. Insurance Data Star Schema and Hierarchy.

Data Model

At a distributed site base data are stored using a STAR schema in a relational database with a central *Fact* table, and associated *Dimension* tables. The database schema for the insurance data is shown in Figure 1 along with the concept hierarchy showing that individual vehicle drivers are nested within *Regions* within *Countries*. The active statistical metadata tables are not defined here. In this particular instance the country level data are stored in each site's metadata store (McClean et al. 2000). Storing the base data in the STAR schema format allows us to leverage the query processing system of SQL databases to efficiently compute the sufficient statistics in the form of the aggregate data. Aggregate data in a MAMED object are stored in a relational table with a schema of the following form:

$$\mathbf{R} <C_1,\ldots C_n; S_1\{N_1, SM_1, SS_1\} \ldots S_m\{N_m, SM_m, SS_m\}>$$

where $C_1,\ldots C_n$ represent n dimension attributes and $S_1,\ldots S_m$ are m sets of summary attributes. Each S_i in the aggregate relation consists of aggregates in the form of count (N), sum (SM) and sums of squares (SS) for a measure or derived measure from the base data. The aggregates in each set of summary attributes are functionally defined by a cross product of the dimension attribute domain values. As an example, Table 1 shows an aggregate relation with two dimension attributes (gender and claim type) and one summary attribute (age). The cross product of the dimension attribute values results in four records, each with its corresponding summary attribute values for the age measure.

Gender	Claim Type	Age_N	Age_SM	Age_SS
Male	Accident	4240	148428	5194534
Male	Theft	1341	45594	1550196
Female	Accident	2210	70720	2263040
Female	Theft	956	26768	749504

Table 1: Aggregate relation for Gender and Claim Type.

When relational data are processed into flat file format, as is most often the case with data mining algorithms, the richness of the structure is often removed (Friedman et al. 1999). This process loses information which might be helpful in understanding the relationships in the data more fully. This is also the case in constructing the aggregate data in a single relation.   When the individual level and higher hierarchy level attributes are integrated into a single aggregate relation for communication to the central site, the hierarchical information is lost. To ensure that this is not the case, the relevant hierarchical information for the attributes is included within the statistical metadata in the system.   This is equivalent to retaining some of the properties of related entities (between a driver entity, a region and a country entity) if the base data which are being aggregated are stored using an entity-relationship model.

Sufficient Statistics

Graefe et al. (Graefe et al 1998) state that "most algorithms are driven by a set of sufficient statistics that are significantly smaller that the data". Their approach takes advantage of the query processing system of SQL databases to produce this set of sufficient statistics for the task of classification, thereby avoiding the need to move the individual level data from the database to the client. This results in a significant increase in performance. In the distributed database situation, where it may be too expensive or prohibited by confidentiality constraints to communicate the individual level data across the network, obtaining such a set of sufficient statistics is even more relevant. However in the distributed database scenario, we must not only find a set of sufficient statistics, but also ensure that they can be combined at the central site in a meaningful way. An important concept which enables us to ultilse the summary attributes in the aggregate relations as sufficient statistics is the additive property of these summary attributes. This property allows us to combine aggregate data from the distributed sites seamlessly. The additive property (Sadreddini et al. 1991) is defined as follows:

$$\sigma (\alpha \text{ UNION } \beta) = \sigma (\alpha) + \sigma (\beta) \qquad (1)$$

where α and β are aggregate relations which are domain compatible (McClean et al. 1998) and $\sigma()$ is an application of a summary attribute function (e.g. SUM) over a measure in α and β. Note for example that if the summary attributes contained average values, average is not an additive summary function and thus such aggregates could not be

combined at the central site. However it is possible to calculate average values using the Sum and Count values in the aggregate relations. Two distributed aggregate relations are domain compatible if they contain the same dimension attributes (each with identical domain sets) and the same measures (each defined on the same units). By the time the MAMED objects are communicated to the central site, MAMED operators have been used to ensure that the aggregate relations are macro compatible before all of the aggregate relations are integrated into the one relational table.

Thus it is possible to combine aggregates over these summary attributes from the distributed sites at a central site for our knowledge discovery purposes, once the data have been suitably harmonised using the statistical meta relations and the MAMED operators.

Discovering Relationships using Aggregate Data

Once the aggregate data have been integrated, the challenge is to use these data rather than the base data to discover interesting relationships. The summary attributes N, SM and SS in the aggregate data can be used as a set of sufficient statistics to compute a large number of statistical procedures including standard deviations, means (Sadreddini et al. 1991) and many linear statistical models. To date with the aggregate data, we have worked with *Multilevel statistical Models* (Goldstein 1995) and *Analysis of Variance* (Anova) models (Neter 1996) in identifying relationships between a single measure attribute and a combination of other attributes at various levels of the hierarchy.

An *Anova* model is used in situations where all of the explanatory attributes of interest are dimensions, some of which are nested within hierarchies (e.g. days at level 1, months at level 2 and years at level 3). No assumption is made in the model about the nature of the statistical relationship between the explanatory attributes and the single measure predictor attribute. More than one hierarchy can be included (for example one hierarchy along the time dimension and another along a geographic dimension) in the model but only one attribute is represented at higher hierarchy levels. Before model computation begins, the CUBE operator (Gray et al. 1996) is applied to the final harmonised, integrated aggregate relation. The set of sufficient statistics for this model now consists of the cubed data in this relation. The statistical metadata contain hierarchical information for the model. An example Anova model represented in equation (2) models *insurance claim costs* using *Gender* (G), *Car-class* (P) and the geographic area hierarchy attributes (*Region* (R) and *Country* (C)). This model also includes interaction terms and nestings within the geographic area hierarchy.

$$COST_{ijk\ln} = \mu + G_i + P_j + C_k + R(C)_{l(k)} + GP_{ij} + GC_{ik}$$
$$+ GR(C)_{il(k)} + PC_{jk} + PR(C)_{jl(k)} + GPC_{ijk} + \varepsilon_{ijk\ln} \quad (2)$$

A multilevel problem is one that concerns the relationships between attributes that are measured at a number of different hierarchical levels. An example is attempting to discover rules and exceptions for the attribute "*Cost of insurance claims*" based on a number of individual level attributes, a number of attributes at a regional level and a further set of attributes at a country level indicated in Table 2. This represents a three-level nested hierarchical situation.

Attribute	Values	Level
COST OF INSURANCE CLAIM	{Continuous}	Individual
CLAIM TYPE	{Accident, Theft}	Individual
GENDER	{Male, Female}	Individual
AGE	{Continuous}	Individual
CAR-CLASS	{A, B, C}	Individual
PREVIOUS CLAIMS	{Yes, No}	Individual
DRIVER TEST SCORE	{Continuous}	Individual
YEARLY MILEAGE	{Continuous}	Individual
REGION NAME		
REGION TYPE	{City, County}	Region
MANDATORY CAR TESTING	{Yes, No}	Region
ROAD ASSESMENT SCORE	{Continuous}	Region
COUNTRY NAME		
DRIVER LOWER AGE LIMIT	{16, 18}	Country
MEAN COUNTRY CLAIM	{Continuous}	Country

Table 2: Attributes of Interest in the Modeling Phase.

Historically such multilevel problems have been analysed by moving all attributes to one single level of interest by aggregation or disaggregation. Aggregation means carrying out the analysis at the region level for example, and using the means of each individual level attribute for each region in place of the individual level data. Disaggregation means moving region level attribute values from being descriptors of regions to being descriptors of the individual.

However these strategies create two different sets of problems. Once groupings exist (e.g. drivers from the same country or members of the same family, properties of the entities we wish to include in the modeling), the individual instances may not be independently and identically distributed (IID). Heterogeneity is very often present in relationships between hierarchical data held in databases, and frequently the statistical models with full IID assumptions do not explain this heterogeneity sufficiently. In fact if this independence assumption is violated, the estimates of the standard errors of conventional statistical tests are much too small, resulting in many deceptive statistically significant results. This lack of independence between observations within groups is expressed as the *intra-class correlation*. It is a population estimate of the variance explained by the grouping structure. Kreft (Kreft 1998) states that the *within-group* information can account for as much as 80-90% of the total variation in the data. Therefore by aggregating to a higher group level this information is lost before the analysis has even begun.

Disaggregation can lead to many apparent statistically significant relationships that are in reality questionable. There can also be problems associated with analysing all of

the data solely at one level and drawing conclusions about the relationships at another level. Robinson (Robinson 1950) shows that aggregate-level relationships (e.g. group means of individual level attributes) can not be used as estimates for the corresponding individual-level relationships. This is known as the Robinson effect or the ecological fallacy. Robinson also showed that drawing inferences at a higher level from an analysis at a lower level can be just as misleading. This is known as the atomistic fallacy.

Multilevel models (Goldstein 1995) attempt more realistically to model these situations where there is clustering of individuals within groups and attributes are measured at different grouping levels of the hierarchy. In these models the assumption of independence of individuals is dropped, and relationships in the data are no longer assumed to be fixed over groups but are allowed to differ. We use multilevel models with a number of goals in mind:

- firstly to determine the direct relationships between individual and group level explanatory attributes (e.g. *age, region type*) and a single measure attribute (*cost of insurance claim*).
- Secondly to determine if the explanatory attributes at the group level (regional and country attributes) serve as moderators of individual level relationships. If such moderators exist, they show up as statistical interactions between explanatory attributes from different levels. This would occur, for example, if varying the *regional road assessment score* affected the relationship between *yearly mileage* and *claim cost* at the individual level. Such *moderator* or *interaction* relationships also allow us to explain some of the variation in the cost of insurance claims between groups. This is not possible using other statistical techniques. As an example, if there is a stark difference in the individual level cost claims between certain regions, it may be possible to explain this difference in the multilevel model using some of the region level attributes (whether the region is a city or a county).
- thirdly to improve prediction of cost claims within individual units (Goldstein 1995), especially for minority groupings by pooling similar individuals from different groups.
- fourthly to partition the variance components among levels into *between-group* (how much of the variation in claim costs can be put down to differences in drivers from different regions) and *within-group* components (how much of this variation is due to differences between drivers within the same region).
- lastly to isolate groups at different levels of the hierarchy that represent exceptions.

The most famous example of how an analysis of multilevel data using standard techniques (in this case multiple regression) can produce incorrect inferences comes from Aitkin et al. (Aitkin et al. 1981). A study had been carried out on different teaching methods for children with teachers in different schools. The results of the analysis using the standard techniques showed that there was a statistical difference between the different teaching methods. As a result the new teaching methods were implemented. However, when the data was analysed using the multilevel models, the results showed that there was in fact no statistical difference between the methods and that the results using the standard statistical methods had been incorrect.

In an insurance scenario, such a "false positive" result may end up costing the company a lot of money if the cost of insurance is incorrectly set according to this apparently significant result.

Model details and Parameter estimation

Parameter estimation for the ANOVA models is a straightforward computational procedure once the relevant aggregate data have been integrated from the distributed sites, integrated, and cubed. The algorithm is not described here and the model is described in (Pairceir 1999). However computing a multilevel model requires an iterative process. The one we use is iterative generalised least squares (IGLS) (Goldstein 1995) as this allows us to use our aggregate data as a sufficient set of statistics

To explain the details of a multilevel model, we take as an example a situation where we are trying to model insurance claim cost (Y) using explanatory attributes *Gender* (X_1) and *DriverTestScore* (X_2) at the individual driver level, and attributes *Region Type* (Z_1) and *RoadAssessmentScore* (Z_2) at the region level. The full equation for this model is shown in (3) below, but it can be broken down for further explanation. This model also incorporates interactions between attributes from different levels in a two level hierarchy.

$$Y_{ij} = [\gamma_{oo} + \gamma_{10}X_{1ij} + \gamma_{20}X_{2ij} + \gamma_{o1}Z_{1j} + \gamma_{o2}Z_{2j} + \gamma_{11}X_{1ij}Z_{1j} + \gamma_{12}X_{1ij}Z_{2j} + \gamma_{21}X_{2ij}Z_{1j} + \gamma_{22}X_{2ij}Z_{2j}] + [\mu_{2j}X_{2ij} + \mu_{1j}X_{1ij} + \mu_{0j} + \varepsilon_{ij}] \qquad (3)$$

Initially a separate model can be built for each region$_j$, with separate slopes and intercepts as shown in (4).

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \varepsilon_{ij} \qquad (4)$$

where $\beta_{0.}$ is the usual regression equation intercept, $\beta_{1.}$ the usual regression slope and ε_{ij} the usual residual error term. Y_{ij} represents the cost of insurance claim for an individual driver$_i$ within region$_j$.

Multilevel models differ from a multiple regression model in that both the intercepts and slopes are allowed to vary across regions. Across all regions, each of these slope and intercept coefficients have a mean and variance, and in the multilevel model the variation in these coefficients is modeled by introducing explanatory attributes at the Region Level (Z_1 and Z_2). Thus for the intercept coefficients across regions, the level 2 model is

$$\beta_{0j} = \gamma_{oo} + \gamma_{o1}Z_{1j} + \gamma_{o2}Z_{2j} + \mu_{0j} \qquad (5)$$

and for each slope coefficient (β_{1j}, β_{2j}) for h$\in$ {1,2}

$$\beta_{hj} = \gamma_{h0} + \gamma_{h1}Z_{1j} + \gamma_{h2}Z_{2j} + \mu_{hj} \qquad (6)$$

where γ_{oo}, γ_{ho} are the intercept coefficients, $\gamma_{o.}$ and $\gamma_{h.}$ the slope coefficients and, μ_{0j} and μ_{hj} the residual errors at the region level.

Equation (5) states that the variation in the level 1 intercept coefficient can possibly be explained by the level 2 attributes Z_1 and Z_2, with any residual error variance being captured in μ_{0j}. Equation (6) states that the variation in the level 1 slope coefficients can possibly be explained by attributes Z_1 and Z_2, with any residual error variance being captured in μ_{hj}. By substituting Equations (5) and (6) into Equation (4) for β_{0j}, β_{1j} and β_{0j}, we obtain a single complex equation giving us a multilevel model at two levels shown in Equation (3). This can easily be generalised to a case with more levels and more explanatory attributes at each level. In equation (3), the square brackets break the model up into a *fixed components* part and a *random* part.

Maximum Likelihood estimation is used to obtain estimates of the fixed and random coefficients of the model. Computing the maximum likelihood estimates requires an iterative procedure, in this case IGLS. Eaach iteration consists of two stages. Both stages involve the use of generalised least squares equations (GLS). The typical form of this equation is shown in equation (7). For the matrix notation in equation (7) below, X is a matrix of explanatory attribute values, V is a Block Diagonal covariance Matrix (Goldstein 1995) containing the variance and covariance components of the model, β is a matrix of model coefficients, and Y is a matrix of the measure attribute values. In equation (7), all values except those in the β (fixed coefficients) and V (random coefficients) matrices are known. If the V matrix values were known, we could obtain a solution to the equation in just a single iteration. However because the random coefficients in the V matrix are unknown, they must also be estimated using GLS.

$$\beta = (X^TV^{-1}X)^{-1} X^TV^{-1}Y \qquad (7)$$

Thus a single iteration moves from using the β matrix coefficients to improve estimates of the random coefficients in a GLS equation similar in form to that of equation (7), to using these improved V matrix random coefficient estimates to obtain improved estimates of the fixed coefficients in β in equation (7). To start the process, initial β matrix coefficients are estimated using ordinary least squares (OLS) equations. The iterations then continue in this way going from one equation to the other until the process converges.

If all of the data is available centrally, a solution is obtained by a number of matrix computations in each iteration. However, in the distributed database case, each iteration requires a number of data communications to and from each distributed database involved. The following algorithm overview illustrates the steps involved in the process. It begins with the user specifying the data of interest and the multilevel model to be used.

Step 1.
MAMED Object request sent from the Query Agent to the relevant distributed sites. The MAMED Objects are computed at each site and returned to the central site where they are harmonised.
Step 2.
The necessary matrices containing aggregates are assembled from the returned harmonised MAMED Object and are used to calculate the initial β model coefficients using ordinary least squares (OLS) equations.
Step 3.
The resulting β model coefficients are communicated to each distributed site and used to compute MAMED Objects consisting of aggregates required to compute the V matrix components.
Step 4.
These new MAMED Objects are returned to the domain site, harmonised, and used to assemble the necessary aggregate matrices required to calculate a V matrix using GLS.
Step 5.
The resulting V^{-1} matrix coefficients are communicated from the domain server to each distributed site and a new MAMED Object incorporating this V^{-1} data is computed for the next iteration at the domain site.[1]
This process then returns to Step 2 and continues in an iterative fashion from refined β estimation to refined V estimation until convergence. After iteration 1, OLS in Step 2 replaced by GLS.

---

[1] Aggregates in the matrices require sums of products of individual values of the **X** matrix and the **V**$^{-1}$ matrix which must be computed at the distributed sites. It is for this reason that the **V**$^{-1}$ matrix must be communicated to the distributed sites.

Relationships, Rules and Exceptions

Once the model coefficients have been computed, we present the significant attribute relationships to the user via a graphical interface rather than the traditional statistical tables. The user may investigate significant relationships at more detailed levels by interacting with the graphical interface. On the basis of these relationships, rules in conjunctive normal form are induced and exceptions to these rules are discovered. Exceptions are groups of individuals at different levels of the hierarchy that represent differences in insurance claim costs relative to other such groups. A group is deemed to be an *exception* if the actual mean insurance claim cost for the group is significantly different in a statistical sense from the value anticipated using the multilevel model. The user may then browse these exceptions at different levels of the hierarchy.
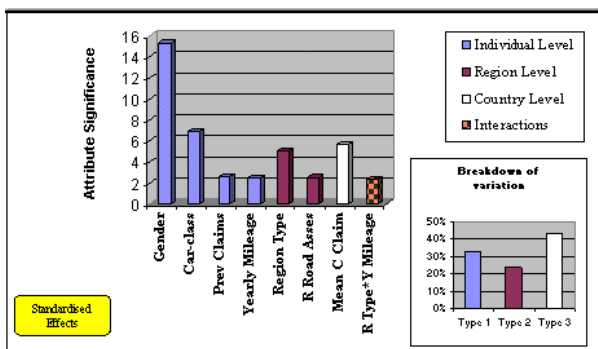

Figure 2. Attribute Significance and Breakdown of Variance between levels.

Figure 2 shows an example of the initial graphical presentation of the model results to the user. In this case it is for a multi-level model which models *insurance cost claims* with exploratory attributes taken from the three levels of the geographic area hierarchy. Our approach summarises the main details of this output in a format more suited to a user not overly familiar with statistical modeling and analysis. Only those explanatory attributes or *effects* from the various hierarchical levels which exhibit a statistically significant relationship with *Cost of Insurance claim* are included in the graph. The higher the bar, the greater the statistical significance of the relationship. There is a significant cross level interaction effect indicated between the region level attribute *region type* and *yearly mileage* from the individual level. This means that the relationship between *yearly mileage* and *insurance claim cost* is also dependent on the associated value of *region type*. The user can interact fully with this graphical output to interrogate the results at more detailed levels. This allows the user to understand an effect's relationship with the cost of insurance claim in greater depth. Dimension attribute-value relationships with insurance cost claims can be viewed graphically in terms of deviations from the overall mean insurance claim cost. The graph in Figure 3 illustrates

this for each car-class in terms of deviance in claim cost from the overall average claim cost.
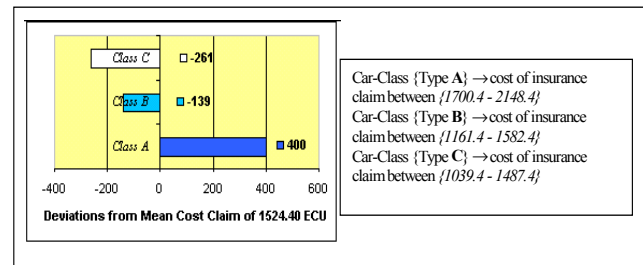

Figure 3. Rules induced from the multilevel model.

As a second step, the significant relationships are turned into rules in conjunctive normal form with associated statistical confidence intervals. Figure 3 contains some example rules relating insurance cost claims with Car-class attribute values. Rules involving measure attributes take on a slightly different format. These rules can also involve interactions between two or more attributes and the insurance claim costs.

Once the rules have been induced, the final step in the results presentation involves the identification of groupings of individual cases which represent exceptions to these rules. These are also presented in conjunctive normal form. An example exception grouping is shown in (3).

[Country {Ireland} ∧ Car-class{A} ∧ Region {City} ∧ Claim Type {Theft} ∧ Gender {Male}, predicted cost claim {1820.54-2023.54 ECU}, actual average cost claim 2100 ECU] (3)

With the multilevel models, it is also possible to present valuable information on the variation components of the data. This information is included in the graphical output in Figure 2. This tells the user that 33% of the overall variation in insurance claim costs was between drivers within regions, 24% was between regions within countries and a large 43% of the variation was between countries. These figures aid the user in determining where most of the variation in the cost of insurance claims arises.

Summary and Related Work

In this paper we have presented our recent and current work in relation to discovering relationships, rules and exceptions from aggregate data retrieved from a number of databases distributed over the internet. We have shown how it is possible to use the aggregate data as a set of sufficient statistics in the building of Multilevel models and Anova models which we use for the discovery purposes. Other distributed data mining work has concentrated on distributing the data to improve the efficiently of data mining algorithms or on using meta-learning techniques to build models (Stolfo et al. 1997, Provost 1998). SarWagi et al. (Sarawagi et al. 1998) worked on the use of Anova

models to highlight exceptions to a user within an OLAP framework.

Future work will concentrate on the application of the distributed aggregate data framework to other knowledge discovery tasks involving hierarchical data. The automation of model building for multi-level models within the framework is also being studied. In this scenario, the user selects a set of attributes of interest and the system would produce the most appropriate model in a step-wise fashion. This may also involve assessing the possible use of random sampling of the distributed databases in building the initial multilevel models. We also intend to look into the possibility of storing previously discovered relationships in the data as part of the textual statistical metadata in XML. This could then be used as part of the domain knowledge. Finally, there is also a need to apply more robust techniques for the production of the final rules.

References

Aitkin M., Anderson, D. and Hinde, J. 1981. Statistical Modelling of data on teaching styles. *Journal of the Royal Statistical Society Series A*. no. 149 1-43.

Friedman, N, Getoor, L., Koller, D. and Pfeffer, A. 1999. Learning Probabilistic Relational Models. *In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, 1300-1309, Stockholm, Sweden.: Morgan Kaufmann.

Goldstein, H. eds. 1995. *Multilevel Statistical Models*. New York.: Halstead Press.

Graefe, G, Fayyad, U., Chaudhuri, S. 1998. On the Efficient Gathering of Sufficient Statistics for Classification from Large SQL Databases. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*. 204-208 California.: AAAI Press.

Gray, J., Bosworth, A., Layman, A., Pirahesh, H. 1996. Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab and Sub-Total. In *Proceedings of the Twelfth International Conference on Data Engineering*, 152-159. Louisiana.: IEEE Computer Society.

Kreft, I.,. de Leeuw, J. eds 1998. *Introducing multilevel modeling*. London, Sage.

McClean S., Grossman, W. and Froeschl, K. 1998. Towards Metadata-Guided Distributed Statistical Processing. In *Proceedings of New Techniques and Technologies in Statistics*. 327-332 Sorrento, Italy:

McClean, S., Páircéir, R., Scotney, B. & Zhang, Y. 2000. Adding context to the retrieval of summary tables from distributed databases via the internet. Submitted to VLDB conference 2000.

Neter, J. 3rd eds. 1996. *Applied linear statistical models*. London.: Irwin.

Páircéir, R., McClean, S., Scotney, B. 1999. Automated Discovery of Rules and Exceptions from Distributed Databases Using Aggregates. *In Proceedings of Third European Conference Principles of Data Mining and Knowledge Discovery*, 156-164, Prague.: LNCS Vol. 1704, Springer.

Provost, F. 1998. Distributed data mining: scaling up and beyond. *KDD workshop on Distributed and Parallel issues in data mining*, KDD-98.

Robinson, W.S. 1950. Ecological correlations and the behavior of individuals. *American Sociological Review*, no. 15, 351-357.

Sadreddini, M., Bell D., and McClean SI. 1991. A Model for integration of Raw Data and Aggregate Views in Heterogeneous Statistical Databases. *Database Technology* vol 4,no. 2: 115-127.

Sarawagi, S., Agrawal, R., Megiddo, N. 1998. Discovery-Driven Exploration of OLAP Data Cubes. In Proceedings of 6th International Conference on Extending Database Technology, 168-182, Valencia, Spain.: LNCS Vol. 1377, Springer.

Shoshani, A. 1997. OLAP and Statistical Databases: Similarities and Differences. In *Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, 185-196, Tucson, Arizona.: ACM Press.

Srikant, R. and Agrawal, R. 1995. Mining Generalised Association Rules. *In Proceedings of 21th International Conference on Very Large Data Bases*, 407-419, Zurich, Switzerland. Morgan Kaufmann.

Stolfo, S., Prodromidis, A., Tselepis, S., Lee, W., Fan, D., Chan P. 1997. JAM: Java Agents for Meta-Learning over Distributed Databases. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*. 74-81California.: AAAI Press.