# Data Mining Audiology Records with the Chi-squared Test and Self-Organising Maps

Michael Oakes, Shaun Cox, and Stefan Wermter

School of Computing and Technology, David Goldman Informatics Centre, Sir Tom Cowie Campus at St. Peter's, University of Sunderland, SR6 0DD, England.
{Michael.Oakes, Shaun.Cox, Stefan.Wermter}@sunderland.ac.uk
http://www.his.sunderland.ac.uk

**Abstract.** In our project AudioMine we wish to address the problem that we need to understand more of the underlying factors influencing which patients would benefit from being fitted with a hearing aid. We describe some results from our pilot study, in which two data mining techniques, the chi-squared test and self-organising maps, were used to discover associations between various fields in 20,000 patient audiology records. We discuss methods of determining the degree of benefit experienced by hearing aid users, so in our main study, working with a larger data set, we will be search for associations between features of audiology records and degree of hearing aid benefit.

## 1 Introduction

A number of authors have worked on the topic of medical data mining using single techniques on homogeneous data. Cios [1] describes the topic in general, while Feldman and Hirsch [2] discuss the data mining of medical texts.

The PROTOS system of Porter, Bareiss and Holte [3] classifies patient audiology records consisting of symptoms, test results and patient history into one of 24 diagnostic categories using Case-Based Reasoning (CBR). Their case descriptions are taken from the DATOS data set [4] which contains 200 records for training and 26 records for testing. As is typical in audiology records, some attributes are missing from the records. Each record has 69 attributes, together with an identifier attribute and a class attribute. Most of the attributes are Boolean, such as age_greater_than_60 or notch_4K which can only take the values true and false. Other attributes can take a range of values, such as the results of speech_audiometry which can be normal, good, very_good, very_poor or unmeasured) or the results of air conduction audiometry which can be mild, moderate, severe or profound hearing losses or normal. The diagnostic categories in the training set include cochlear_age, cochlear_age_and_noise, possible_menieres, and normal_ear.

PROTOS performs a classification by finding the class (diagnostic category) exemplar which most closely matches each new case, and explains the classification. The system employs supervised learning, since if the domain expert disagrees with either the classification or the explanation, the exemplars for the category must be adjusted

until the expert is satisfied with both the classification and the explanation. In the PROTOS system categories can be represented by more than one exemplar. This is done for any category which is polymorphous, where there is unexplained variability between the members of that category. If the expert wants to place a case in a category which differs significantly from the existing exemplars of that category, this case becomes an additional exemplar for that category. The degree of match between cases and exemplars can be increased or decreased when the expert specifies relations of varying strengths between attributes and categories, as in the following examples:

```
  Speech(poor)      is      usually      equivalent      to
speech(very_poor).
  Bone(unmeasured)     is     sometimes     equivalent     to
bone(abnormal).
  Acoustic_reflex_threshold(elevated)   is   spurious   to
age_and_noise_induced_cochlear.
```

Other authors have used the DATOS data set, but their main interest was in the comparison of supervised models for classification, rather than on the audiology data set per se. Holmes and Trigg [5] used DATOS among other data sets with their method of comparing tree-based supervised classification algorithms, which was based on approximate tree matching algorithms. Dietterich [6] also worked on the construction of decision trees, while Chickering and Heckerman [7] used DATOS as one of their real world data sets to compare various Bayesian models.

Discussions of how audiological data, medical reports and the hearing aid dispenser's judgement are presently combined for the selection of hearing aid users and their hearing aids is given by Nunez, Clarke, Hawthorne and Robertshaw [8], and more recently by Arlinger, Lyregaard, Billeermark and Oberg [9]. Other authors discuss hearing aid outcomes in groups of hearing impaired patients, such as older patients [10] [11] or those with conductive losses [12]. Walker shows that one factor influencing hearing aid outcomes is the frequency response of the hearing aid [12]. Anari, Axelsson, Eliason and Magnusson [13] describe the combination of audiological data (results of hearing tests and questionnaires) to classify patients suffering from hypersensitivity to sound.

## 2 The Data Set Used in the Pilot Study

For our pilot study we were granted access to a very large electronic database of audiological data, consisting of 180,000 individual records covering 23,000 different patients, stored in a relational database system at James Cook University Hospital, Middlesbrough. The results reported in this paper were produced using a subset of over 20,000 of these records. The data stored within each record is heterogeneous, and consists of the following elements:

1) Audiograms, graphs showing an individual hearing threshold (the faintest sound he or she can hear) in each ear, typically at six different pitches. Two graphs are

obtained for each ear, one by air conduction (using sounds from a headphone on the ear, measuring overall hearing ability), and one by bone conduction. The sound is presented to the mastoid bone behind the ear, measuring the hearing ability of the inner ear (cochlea and auditory nerve).

2) Structured tabular data (fields for hearing aid type, date of birth, etc. as in a conventional database), e.g. `|BE19|`, `|28-05-1921|`.
3) Unstructured text as phrases or short sentences, e.g. `|MOULD SCRATCHING|` (specific observations made about each patient's case).

The records are in the form of a relational database consisting of six tables. One of the tables, *audiolog,* contains records in the following format, with fields delimited by vertical bars:

```
062D726|30|30|55|50|45|80|30|45|50|60|50|90|5|20|45|55|
30||||||28-06-1991|JILLIAN|
```

The first field is the patient number, followed by six fields for right ear air conduction, six for left ear air conduction, six for right ear bone conduction, six for left ear bone conduction (not filled), and finally the date of the hearing test and the name of the audiologist.

To some extent this audiology record can be seen as a scaled-down version of a medical record since the heterogeneous character of the audiology records is representative for medical records in general. In almost all cases there is a combination of structured and unstructured information which makes it difficult to process, for instance with direct queries from a relational database system. The data mining techniques described in this paper are designed to take advantage of the complementary nature of these three different types of information.

## 3 No Learning: Basic Statistical Chi-Squared Tests on Categorial Representations

The chi-squared test is used to determine whether two events occur together more often than one would expect by chance. It is designed for work with nominal (also called categorial) data, such as the attributes found in tabular data. Nominal facts are data that can be sorted into categories such as the diagnostic category of a patient. A binary decision is made whereby a patient either does or does not belong to a given category, and no category has precedence over any other.

Zembowicz and Zytkov [14] describe the simple but elegant "49er" technique that uses the chi-squared test to scan the fields of a database to find which pairs of attributes tend to occur together. For example, one might wish to determine the fact whether characteristic A tends to co-occur with characteristic B. By considering every record in the database four combinations of events are found: a) A and B occur to-

gether; b) A occurs but B does not; c) A does not occur but B does occur; d) neither A nor B occur. Chi-squared ( $X^2$ ) is then calculated using the following formula:

$$X^2 = N ( | ad – bc | - N / 2)^2 \ / \ ( a + b )( c + d )(a + c )( b + d ) . \quad \textbf{(1)}$$

where N = a + b + c + d, i.e. the total number of records in the database. If chi-squared is greater than 3.84 we can be 95% confident that A and B really do occur together more often than one would expect by chance; and if chi-squared is more than 6.64 there is 99% confidence. This method can be extended for all three types of data, and for finding relationships between them. Consider these four hypothetical data items:

1) Sex is female, an example of the tabular data in the record.
2) The air conduction (overall hearing) threshold is 40 dB (decibels), as registered on the audiogram.
3) The accompanying text contains the word “otosclerosis”.
4) The hearing aid fitting was successful, as determined by battery usage or frequency of repairs.

To find whether gender and successful hearing aid usage tend to go together, one should count a) the number of records showing both female gender and successful hearing aid use, b) the number of records where gender is female but hearing aid use was not successful, c) the number of cases where gender was not female but hearing aid use was successful, and d) the number of cases where gender was not female and hearing aid use was not successful.

Although the audiogram displays numeric data, the points on the graph are only plotted at discrete 5dB intervals and typically at six frequencies. Bands of thresholds can be grouped into larger nominal categories, such as threshold at or above 40 dB, and those below 40 dB. Then the same four combinations can be examined. Finally, with regard to the textual data, one can for example determine whether the presence of a word tends to occur with successful hearing aid outcomes. This should be done with every non-stoplisted mid frequency word, and for sequences of words. This technique is a synthesis of Zembowicz and Zytkov’s 49er technique and Rayson, Leech and Hodgson’s method of determining whether certain vocabulary is typical of different social groups [15].

When examining the associations between gender and diagnostic category, we found that that tinnitus (ringing in the ear) was present in 632 men and 541 women, showing that this was more prevalent in men ($X^2 = 16.4$). However, there were no significant associations between types of tinnitus maskers and gender. Tables 1 and 2 show all the significant associations found between hearing aid mould type and gender. Associations between hearing aid type and gender are reported in Cox et al. [16].

**Table 1.** Associations between gender and hearing aid mould type (left ear)

| Mould Type | Female | Male | $X^2$ |
|---|---|---|---|
| 2107V2 | 101 | 133 | 7.2 |
| BPORE | 29 | 10 | 7.9 |
| IROS | 24 | 95 | 48.1 |
| N1 | 18 | 3 | 9.6 |
| N2 | 0 | 8 | 8.6 |
| N8 | 139 | 265 | 50.2 |
| V2 | 404 | 762 | 145.7 |

**Table 2.** Associations between gender and hearing aid mould type (right ear)

| Mould Type | Female | Male | $X^2$ |
|---|---|---|---|
| 2107V1 | 912 | 677 | 20.6 |
| BPORE | 32 | 13 | 6.6 |
| IROS | 24 | 80 | 34.7 |
| N1 | 27 | 6 | 11.8 |
| N8 | 141 | 253 | 41.5 |
| V1 | 2160 | 1776 | 16.7 |
| V2 | 485 | 792 | 104.4 |

## 4 Unsupervised Learning: Self-Organising Maps

The chi-squared test is attractive based on its simplicity and therefore speed efficiency on larger data mining experiments with many comparing evaluations. While the chi-squared technique can reveal associations between pairs of variables, multivariate neural techniques can examine interactions between greater numbers of variables.

Unsupervised learning of multivariate representations can be performed in self-organising maps (SOM). SOMs shall be used to cluster and classify unstructured and structured portions of the audiology records. The SOMs complement the supervised networks in that it uses competitive learning and as such is an unsupervised technique [17][18]. Each node in the map contains a model vector containing the same number of elements as the input vector. In the example of Figure 1, the vectors contained 6 elements, corresponding to air conduction thresholds at 250, 500, 1000, 2000, 4000 and 8000 Hz. Each new input vector is compared with all the model vectors. The node whose model vector best matches the input vector is then identified and referred to as the winner. The model vectors of the winner and a number of its neighbouring nodes

are changed towards the input vector. In this way, a high-dimensional feature space can be clustered and analysed in a 2- dimensional or 3-dimensional map.

In the SOM shown in Figure 1, produced using the MATLAB programming environment, various diagnostic categories are clustered according to the patient's air conduction results. The clusters are represented by the corners of the triangles, and the number of records with each diagnosis allocated to that cluster is written to the left of the cluster point. All the clusters contain records for tinnitus sufferers, showing that tinnitus can occur with a wide variety of air conduction test results. However, all the meningitis sufferers are allocated to clusters near the top right of the SOM, showing that they have similar air conduction results. Most of the patients with conductive losses, such as otosclerosis, glue ear, cleft palate and otitis externa, cluster near the top left of the SOM. A SOM showing how various types of hearing aids cluster according to air conduction thresholds is given in [16].

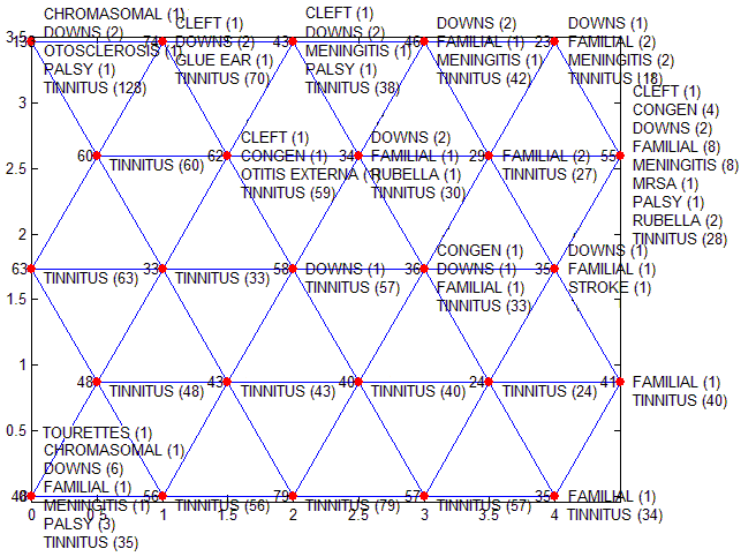


**Fig. 1.** A self-organising map displaying diagnosis clustered according to air conduction thresholds at six frequencies

## 5 Calibration of the Study

In our pilot study we noted a number of interesting correlations between the attributes of our database, but there was no attribute for the success or otherwise of the hearing aid fitting. Since the prediction of the hearing aid prognosis from the other

audiology records is our ultimate aim, we need to be able to substantiate the degree of benefit experienced by each hearing aid user, and hence calibrate our study. Assessing the degree of usage, benefit and user satisfaction of a hearing aid is notoriously difficult [19], but non-usage of a hearing aid by a patient is more likely if the patient never returns to the clinic for routine hearing aid maintenance. We can be very confident that there is a good correlation between regular returns to the clinic for hearing aid maintenance and hearing aid usage. A hearing aid breaks down on average about once per year, and it is very rare for a hearing to work without any maintenance for more than two years.

A survey of the literature reveals three other main criteria for estimating hearing aid benefit: a) speech audiometry in noisy conditions (SIN) [20]; b) self report of hours of hearing aid use [11]; c) subjective appraisals by hearing aid users, such as the Client-Oriented Scale of Improvement (COSI) [21], the Satisfaction Questionnaire [22] and the Glasgow Hearing Aid Benefit Profile (GHABP) [23]. Although SIN is time consuming, and the speech materials and laboratory test conditions have not yet been standardised [24], SIN testing will be available at James Cook Hospital for a limited number of patients, and has the advantage that it is the most objective measure of hearing aid benefit.

Through the NHS-funded Modernising Hearing Aid Services (MHAS) network we have been now given access to an even larger electronic database of audiological data, consisting of over 300,000 individual records, also stored at the James Cook Hospital in Middlesbrough.

## 6 Conclusion: Reasons for a Data Mining Approach

A data mining approach is required for our forthcoming main study, since no individual would be able to analyse 300,000 individual records. In our pilot study, we have shown that data mining is able to discover interesting relationships between the data. Cios and Moore [25] state that "modern medicine generates huge amounts of data, but at the same time there is an acute and widening gap between data collection and data comprehension". Data mining in medicine is characterised by the processing of heterogeneous data, typically including (as in our case) the unstructured English characteristic of audiology records. Our approach considers structured fields, text tokens and audiograms.

## References


1. Cios, K. Medical Data Mining and Knowledge Discovery. Springer Verlag, Berlin Heideberg New York (2001)
2. Feldman, R. Hirsh, H. Finding Associations in Collections of Text. In Michalski, R., Bratko, I., Kubat, M. editors, Machine Learning and Data Mining, John Wiley, Chichester (1998) 223-240
3. Porter, B., Bareiss, R., Holte, R. Concept Learning and Heuristic Classification in Weak-Theory Domains. Artificial Intelligence 45 (1990) 229-263

4. The DATOS data set. http://ccc.inaoep.mx/DATOS/audiology
5. Holmes, G., Trigg, L. A Diagnostic Tool for Tree-based Supervised Classification Learning Algorithms. Working Paper 99/3, University of Waikato (1999)
6. Dietterich, T. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting and Randomisation. Machine Learning 40 (2000) 1
7. Chickering, D., Heckerman, D., Efficient Approximations for the Marginal Likelihood of Bayesian Networks with Hidden Variables, Technical Report MSR-TR-96-08, Microsoft Research, Advanced Technology Division (1996, revised 1997)
8. Nunez, D., Clarke, G., Hawthorne, M., Robertshaw, D. Hearing Aids - Is ENT Evaluation Necessary? British Journal of Audiology (1991)
9. Arlinger, S., Lyregaard, P., Billermark, E., Oberg, M. Fitting Hearing Aids to First Time Users. Scandinavian Audiology 29 (2000) 150-158
10. Hawthorne, M., Nunez, D., Clarke, G., Robertshaw, D. Direct Referral Hearing Aid Provision in the Over Sixties Age Group. Journal of Laryngology and Otology. 105 (1991) 825 – 827
11. Hickson, M., Trimm., M., Worrall, L. Bishop, K. Hearing Aid Fitting: Outcomes for Older Adults, Australian Journal of Audiology 21 (1999)
12. Walker, G., The Required Frequency Responses of Hearing aids for People with Conductive Hearing Losses, Australian Journal of Audiology, 21 (1999)
13. Anari, M., Axelsson, A., Eliason, A., Magnusson, L. Hypersensitivity to Sound – Questionnaire Data, Audiometry and Classification. Scandinavian Audiology, 28 (1999) 219-238
14. Zembowicz, R., Zytkov, J. From Contingency Tables to Various Forms of Knowledge in Databases. In Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. editors, Advances in Knowledge Discovery and Data Mining, AIII Press/MIT Press Cambridge Massachusetts (1996) 329-349
15. Rayson, P., Leech, G. Hodges, M. Social Differentiation in the Use of English Vocabulary. International Journal of Corpus Linguistics 2 (1997) 133-152
16. Cox, S., Oakes, M., Wermter, S. Audiomine: Medical Data Mining in Heterogeneous Audiology Records. International Journal of Computational Intelligence 1 (2004) 1-12
17. Kohonen, T. Exploration of Very Large Databases by Self Organising Maps. In Proceedings of the International Conference on Neural Networks (ICNN97) IEEE Service Centre Piscataway NJ USA (1997) 1-6
18. Honkela, T. Self Organizing Maps in Symbol Processing. Springer-Verlag, Berlin Heidelberg New York (2000)
19. Paul, R., Cox, R. Measuring Hearing Aid Benefit with the APHAB: Is this as Good as it Gets? American Journal of Audiology 4 (1995)
20. Killion, M. The SIN Report: Circuits Haven't Solved the Hearing in Noise Problem. The Hearing Journal, 50 (1997) 28-34
21. Dillon, H., James, A. Ginis, J. Client Oriented Scale of Improvement (COSI) and its Relationship to Several other Measures of Benefit and Satisfaction Provided by Hearing Aids. Journal of the American Academy of Audiology, 8 (1997) 27-43
22. Humes, L., Halling, D., Coughlin, M. Reliability and Stability of Various Hearing Aid Outcome Measures in a Group of Elderly Hearing Aid Wearers. Journal of Speech and Hearing Research, 39 (1996) 923-935
23. The Glasgow Hearing Aid Benefit Profile. www.ihr.gla.ac.uk/products/ghabp.php
24. Walden, B., Towards a Model Clinical-Trials Protocol for Substantiating Hearing Aid User-Benefit Claims, American Journal of Audiology, 6 (1997)
25. Cios, K. Moore, G. Medical Data Mining and Knowledge Discovery: Overview of Key Issues. In Cios, K., editor, Medical Data Mining and Knowledge Discovery, Springer Verlag, Berlin Heidelberg New York (2001)