World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:1, No:1, 2007

AudioMine: Medical Data Mining in Heterogeneous Audiology Records

Shaun Cox, Michael Oakes, Stefan Wermter and Maurice Hawthorne

***Abstract*—** We report on the results of a pilot study in which a data-mining tool was developed for mining audiology records. The records were heterogeneous in that they contained numeric, category and textual data. The tools developed are designed to observe associations between any field in the records and any other field. The techniques employed were the statistical chi-squared test, and the use of self-organizing maps, an unsupervised neural learning approach.

***Keywords*—**Audiology, Data Mining, Chi-squared, Self Organizing Maps.

I. Introduction

In our project AudioMine we 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. Through a variety of data mining techniques, we aim to discover and examine factors influencing the success of a hearing aid fitting.

We have access to a very large 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 data stored within each record is heterogeneous, consisting of an audiogram (graph of hearing ability at different frequencies), structured data (address, data of birth, etc) and unstructured data in the form of free text (specific observations made about each patient's case).

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. We use both statistical and neural techniques for processing this heterogeneous form of medical records in order to combine their various strengths. This use of these techniques is novel for audiology records and has not been performed before.

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 Kononenko, Bratko & Kukar [2] examine data mining for medical diagnosis. The PROTOS system of Porter and Bareiss [3], which employs case based reasoning, was designed to perform heuristic classification tasks in the domain of audiology. Other authors have performed data mining on various data sets, of which audiological data was just one source [4][5]. The NOAH system is an environment for large databases with some data mining options. This system has been used for audiology records [6]. Palisades Research have developed an audiometric analysis system for evaluating hearing conservation programmes. Most such tools are based on traditional generic software components which are not particularly developed for audiology such as SAS or SPSS. In spite of this recent work on medical records and audiology, there is currently no hybrid knowledge discovery tool available which can deal with the combination of structured personal information, audiograms and unstructured phrases as found in our audiology records.

2. THE DATA SET

We have been given access to a large database containing 180,000 individual audiology records covering 23,000 patients, stored in a relational database system at the James Cook University Hospital in Middlesbrough. These records contain three different kinds of data:

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, where 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).
3) Unstructured text (phrases or short sentences).

Manuscript received November 30, 2004.
S. Cox is a student completing his M.Sc. programme in Intelligent Systems at the School of Computing and Technology, University of Sunderland, England; (e-mail: shaun.cox@sunderland.ac.uk)
M. P. Oakes is a Senior Lecturer at the School of Computing and Technology, University of Sunderland, England. (e-mail: michael.oakes@sunderland.ac.uk).
S. Wermter is Professor in Hybrid Intelligent Systems at the School of Computing and Technology, University of Sunderland, England; (e-mail: stefan.wermter@sunderland.ac.uk).
M. Hawthorne is an Ear, Nose and Throat Consultant at James Cook University Hospital, Middlesbrough, England.

II. METHODOLOGY

A. Chi-Squared Test

Our first approach was to use a basic statistical test which did not involve any machine learning. The chi-squared test is a statistical test 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 [7]. Nominal facts are data that can be sorted into categories such as the grammatical category of a word. A binary decision is made whereby a word either does or does not belong to a given category, and no consideration is taken, for example, of whether any grammatical category has preference over any other, or of the degree to which a word fits into a category. Only the overall number of words which fit into each category is considered.

Zembowicz and Zytkov [8] describe the simple but elegant "49er" technique which 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 whether characteristic A tends to co-occur with characteristic B. By consideration of every record in the database, the number of instances in which the following four combinations of events are found: a) A and B occur together; b) A occurs but B does not; c) A does not occur but B does occur; d) neither A nor B occur. These four quantities are traditionally written in a 2 x 2 contingency table. Chi-squared (X^2) is then calculated using the following formula:

$$X^2 = N (| ad – bc | - N / 2)^2 \ / \ (a + b)(c + d)(a + c)(b + d)$$

where N = a + b + c + d, i.e. the total number of records in the database. If chi-squared is greater than 3.84 there is 95% confidence 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 to cater 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 such factors as battery usage or frequency of repairs.

To find whether gender and successful hearing aid usage tend to go together, 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. Thus by not considering, for example, that 60 dB is louder than 30dB, but merely regarding thresholds of 60dB and 30dB as separate categories, the data can be regarded as nominal. 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 possibly for sequences of words. This technique is a synthesis of Zembowicz & Zytkov's 49er technique and Rayson, Leech and Hodgson's [9] method of determining whether certain vocabulary is typical of certain social groups such as men and women (1997). Oakes et al. [10] successfully used the chi-square technique for keywords in texts about pharmacology to classify those texts according to subtopic.

Many separate evaluations must be performed: one for each possible value of every field of all three data sets. Our eventual goal is to seek to find which values of which attributes produce the highest values of chi-square values when compared with successful and unsuccessful hearing aid use, and hence act as predictors of future hearing aid success or otherwise.

B. Self-Organizing 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, it is intended to extend this work by the use of multivariate techniques which examine interactions between greater numbers of variables. One way to do this is to use unsupervised learning in neural networks, for example Self-Organizing Maps (SOM).

Learning techniques as in neural networks have the ability to learn a flat analysis during medical mining in a robust manner [11] and this is the motivation for our decision to explore neural networks for data mining. Unsupervised learning of multivariate representations can be performed in self-organizing maps (SOM), originally developed by Kohonen [12]. Patterson [13] defines the SOM as a "*competitive, self – organizing neural network which learns from its environment without the aid of a teacher*" and as such, is able to both classify input vectors according to the manner in which they are grouped within input space whilst learning their distribution - meaning that neurons next to one another will respond to similar input vectors.

Architecturally the SOM appears deceptively simple, as shown in Fig. 1, possessing in the majority of cases a set number of input units corresponding to the dimensionality of its training vectors, which unlike other ANN serve only to distribute each input vector to the networks output neurons; commonly they are arranged as a 2D grid although deviations on this arrangement do exist. It is these output neurons which are subsequently used to cluster each set of input vectors.

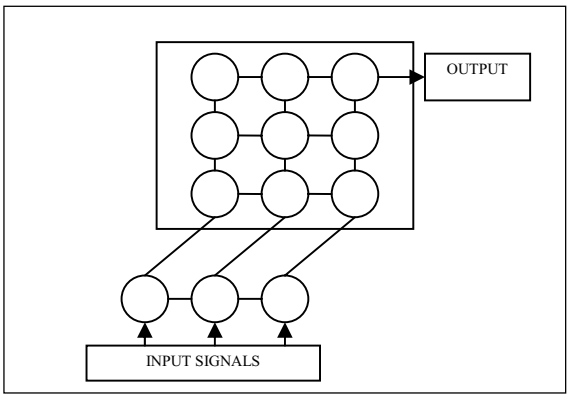

FIG.1
ARCHITECTURE OF A SOM

Although not shown in Fig. 1, each input neuron is in fact connected to every output neuron. The SOM's learning process can be summarized as follows: Learning begins usually with the random initialization of weights ($\mathbf{w}_r$) (the usual strategy employed here is through the use of small random numbers) between the input layer and output layer as well topographical neighborhood parameters and a learning rate ($\boldsymbol{\alpha}$) is set usually at a relatively high value of 0.5. A vector denoted by $\boldsymbol{x}$ is then chosen from the input patterns for input to the network. The output layer neuron **r** with the lowest Euclidean distance i.e. the unit closest to the original input vector $\boldsymbol{x}$ is determined by computing:

$$\| \mathbf{w_r} - \boldsymbol{x} \| = \mathbf{min_r} \{ \| \mathbf{w_r} - \boldsymbol{x} \| \}$$

Weight vectors on the $(t + 1)$th iteration are updated according to

$$\mathbf{w_r}(t + 1) = \mathbf{w_r}(t) + \alpha(\boldsymbol{x} - \mathbf{w_r}(t)) \text{ for units } \mathbf{r} \in N\mathbf{r}$$

$$\mathbf{w_r}(t + 1) = \mathbf{w_r}(t) \text{ for units } \mathbf{r} \notin N\mathbf{r}$$

where $N\mathbf{r}$ is the neighborhood of output neuron **r.** Note that only those weights connected to the winning neuron and its neighbors are updated. This differs from what takes place during learned vector quantization. Finally neighborhood and learning rate parameters are reduced at each epoch [13].

III. RESULTS

In one experiment, we examined the association between the type of hearing aid worn and the patient's age. The results for the left ear are shown in Table I.

TABLE I
SIGNIFICANT ASSOCIATIONS BETWEEN LEFT HEARING AID USAGE AND PATIENT AGE

Aid Type	Age < 16	Age 16-40	Age 41-65	Age > 65	Chi-Squared
ITENN	13	57	246	1145	7.84
BE34	5	77	211	1346	15.44
BE19	15	39	164	1194	18.09
BE37	2	3	2	27	10.56
ITEHH	11	36	173	649	22.55
PPCL	2	32	13	118	101.40
BE18	6	35	69	471	8.14
ITEHN	24	95	401	1895	10.49
BE14	1	11	5	39	35.29

For three degrees of freedom, a chi-squared value > 7.82 shows significance at $p < 0.05$, and a chi-squared value > 11.35 shows significance at $p < 0.01$. Given that the total number of left ear hearing aids prescribed for the age groups in ascending order was 127, 576, 1867 and 10184, PPCL aids were proportionally more often prescribed to younger patients. For the right ear, significant associations between hearing aid type and gender were found for BE19, ITEHN, CI, PPCL and ITEHH, all with $p < 0.01$. Once again the strongest association was found for PPCL (Chi-squared = 94.16), which was more often prescribed to younger patients. Hearing aids which are worn inside the ear have all have the prefix ITE, but different codes for amplifier type and microphone response.

TABLE II
SIGNIFICANT ASSOCIATIONS BETWEEN LEFT HEARING AID USE AND GENDER

Aid Type	Male	Female	Chi-Sq.
ITENN	657	811	7.04
BE34	746	897	5.16
BE19	744	633	13.59
BE51	22	48	7.82
ITEHH	691	189	338.19
ITENH	735	256	280.90
ITEHN	1094	1342	11.10
BE11	102	67	10.21
BE201	54	92	7.28
ITENL2	20	53	12.58
BE104	80	120	5.31
ITENL	85	170	22.54
ITEKH	17	5	7.50
BE102	14	5	4.98

Experiments were also performed comparing the type of hearing aid used with patient gender. The results for the left ear are shown in Table II. For three degrees of freedom, a chi-squared value > 3.841 shows significance at $p < 0.05$, and a chi-squared value > 6.635 shows significance at $p < 0.01$. The strongest associations were that ITEHH and ITENH are

favored for men, a result also seen when repeating the experiment for the right ear: Chi-squared for ITEHH was 295.79, and Chi-squared for ITENH was 224.91. The results given here are just examples of the types of associations the AudioMine tool can discover. It has been designed so that any field of the patient audiology records can be compared against any other. To illustrate the process of data mining audiology records with a self-organizing map, a 5 x 5 self – organizing map is shown to display the results of correlating air – conduction hearing aid thresholds for the left ear with the specific type of hearing aid worn on that ear. The results in Fig. 2 show that clusters of patients with similar degrees of hearing loss tend to wear hearing aids which have similar properties of acoustic gain. The values in parentheses after each hearing aid type show the number of patients wearing this type of hearing aid clustering at each node. Hearing aids with the prefix BE are worn behind the ear, and those with prefix ITE are worn inside the ear. All hearing aids with the prefix BE1 have similar acoustic properties, the fourth digit denoting the manufacturer. Similarly all hearing aids with the prefix BE3 have similar acoustic properties to each other, being more powerful than the BE1 series. Fig. 2 was produced using software written by Simon Lundell [14].

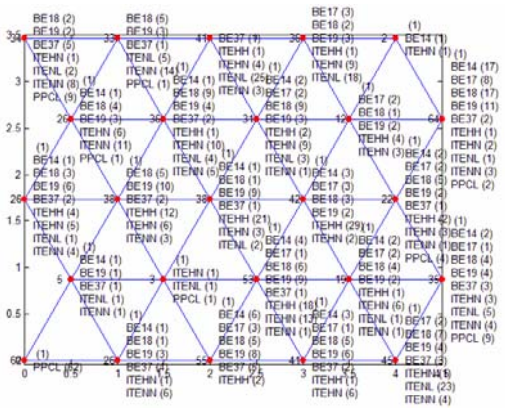

FIG. 2
OUTPUT OF A SOM WHERE HEARING AIDS ARE CLUSTERED ACCORDING TO THE WEARER'S AIR CONDUCTION THRESHOLD FOR THE LEFT EAR.

IV. Conclusion

We have completed a pilot study where we have demonstrated the feasibility of mining for associations in audiological records. James Cook University Hospital have joined a national initiative called Modernising Hearing Aid Services (MHAS), which requires that all data collected at the audiology clinic be stored in an Auditbase database. We plan to adapt our AudioMine tools to cater for the records of MHAS in further extensive work. The modified tool will be suitable for the data held at most other hospitals in the UK, enabling us to conduct data mining surveys on a national scale. As part of the MHAS initiative, hearing aid users will respond to an online survey of satisfaction with their hearing aids in various everyday situations. The availability of this data will open up new data mining opportunities, and help us in our ultimate goal of predicting the success or otherwise of a hearing aid fitting.

Acknowledgment

We wish to thank Graham Clarke and Martin Sandford at the Ear, Nose and Throat Clinic at James Cook University Hospital in Middlesbrough, England, for making the large set of audiology records available to us.

References

[1] K. Cios. *Medical Data Mining and Knowledge Discovery*. Berlin/Heidelberg: Springer Verlag, 2001.
[2] I. Kononenko, I. Bratko and M. Kukar. "Application of Machine Learning to Medical Diagnosis". In R. Michalski, I. Bratko and M. Kubat (editors), *Machine Learning and DataMining*, Chichester: John Wiley, 1998, pp389-408.
[3] B. Porter and E. Bareiss. "PROTOS: An experiment in knowledge acquisition for heuristic classification tasks". *Proc. First International Meeting on Advances in Learning (IMAL)* Les Arcs, France, 1986, pp. 159-174.
[4] G. Holmes and L.Trigg. A diagnostic tool for tree-based supervised classification learning algorithms. URL: www.cs.waikato.ac.nz/~ml/publications/1999/99GH-LT-Diagnostic-Tool.pdf
[5] T. Dietterich. "An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting and randomization". *Machine Learning* 40(2), 2000.
[6] B. Ingrao, "Portable electronic patient records". *Hearing Review* 8(6), 2001.
[7] M. Oakes, *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press, 1998.
[8] R. Zembowicz and J.Zytkov, "From contingency tables to various forms of knowledge in databases.", In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy (editors), *Advances in Knowledge Discovery and Data Mining,* , Cambridge Massachusetts: AIII Press/MIT Press, 1996, pp. 329-349.
[9] P. Rayson, G.Leech and M. Hodges. "Social differentiation in the use of English vocabulary." *International Journal of Corpus Linguistics* 2(1), 1997, pp. 133-152.
[10] M. Oakes, R Gaizauskas, H Fowkes, et al. "Comparision between a method based on the chi-square test and a support vector machine for document classification", *Proceedings of ACM-SIGIR*, New Orleans, 2001.
[11] S. Wermter and V. Weber. "SCREEN: Learning a flat syntactic and semantic spoken language analysis using artificial neural networks". *Journal of Artificial Intelligence Research*, 6(1):35--85, 1997.
[12] T. Kohonen. "Self organisation of very large documents: State of the art." *Proc. ICANN98, the 8th International Conference on Artificial Neural Networks*, volume 1, 1998. London: Springer 1998. pp. 65-74.
[13] D.W. Patterson. *Artificial Neural Networks: Theory and Applications.* Singapore: Simon & Schuster/Prentice Hall, 1996.
[14] S. Lundell, 2001. http://www.ida.his.se/ida/kurser/ai-ann/kursmaterial/tutorial/node38.html.