# Clustering Audiology Data

**Muhammad Naveed Anwar** **naveed.anwar@sunderland.ac.uk**
**Michael P. Oakes** **michael.oakes@sunderland.ac.uk**
**Stefan Wermter** **wermter@informatik.uni-hamburg.de**
**Stefan Heinrich** **heinrich@informatik.uni-hamburg.de**

**Department of Computing, Engineering & Technology, University of Sunderland, Sunderland, UK**

**Department of Informatics, University of Hamburg, Hamburg, Germany**

**Keywords:** audiology, k-means, chi-squared, SOM.

## Abstract

In this paper we describe new results of statistical and neural data mining of audiology patient records, with the ultimate aim of looking for factors influencing which patients would most benefit from being fitted with a hearing aid. We describe how a combination of neural and statistical techniques can usefully subdivide a set of patients into clusters, based on their hearing thresholds at six different frequencies, and then label the clusters with meaningful text labels. In our first experiment we cluster the patients based on similarities between their audiograms using k-means clustering, resulting in two main clusters. We then use the chi-squared test to label each cluster with the keywords selected from the text comment, diagnosis and hearing aid type associated with each patient which are most typical (and atypical) of each cluster. In our second experiment, we again cluster the patients based on similarities between their audiograms, but this time using a self-organizing map (SOM). Here the locations in the resulting map, corresponding to individual patients, are labeled with the type of hearing aid selected for each patient. We demonstrate that this automatic textual labeling addresses well the heterogeneous character of medical audiology records, since they consist of numeric, structured and free text data.

## 1. Introduction

In medicine a substantial amount of new data is created constantly and the amount of data produced is much higher than the amount of knowledge produced. Therefore, large increases in the production of data require a quick transfer to knowledge. To achieve this, clustering examines groups of data items that are similar and dissimilar. It identifies areas of high sample density (data clusters) and shows the centers of these clusters (Principe, Euliano, & Lefebvre, 2000). Clustering techniques include statistical and artificial neural network approaches.

In unsupervised clustering unlabelled data is collected without any supervised knowledge (Wermter, & Sun, 2000). In this paper, we will present an approach of integrating unsupervised clustering of patient audiology data with the identification of textual keywords associated with each cluster, in particular those related to text comment, diagnosis and hearing aid type. Previous authors have made use of the TF.IDF measure, widely used in search engine technology for automatic indexing, for finding the words most associated with clusters. For instance, Maqbool and Babri (2006) used the TF.IDF (Term Frequency Inverse Document Frequency) measure to identify the words extracted from comments in computer code which best characterized clusters of code found by hierarchical agglomerative clustering. Similarly, Ke et al. (2005) selected words from email messages using TF.IDF to characterise clusters of emails found by k-means clustering. The TF.IDF weighting W for a particular label i (such as a word) with respect to a particular cluster j is given by the formula

$$W_{ij} = TF_{ij}.\log_2\left(\frac{N}{NDoc_i}\right)$$

where $TF_{ij}$ is the number of times word i is seen in cluster j, N is the total number of clusters, and

Appearing in *Proceedings of the 19$^{th}$ Machine Learning conference of Belgium and The Netherlands.* Copyright 2010 by the author(s)/owner(s).

$NDoc_i$ is the number of clusters which contain word i. In this paper, we use an alternative technique based on the chi-squared measure for finding the vocabulary associated with clusters, which is better suited for labeling small numbers of clusters (Manning, Raghavan, & Schütze, 2009).

## 2. Audiology Data Repository

For this study, we obtained audiology data from the James Cook University Hospital, Middlesbrough, England. The audiology database contains 180,000 individual records covering 23,000 different patients. It contains heterogeneous records which consist of

- Audiograms, which are the graphs of hearing ability at different frequencies in each ear. Two graphs (AC and BC) are obtained for each ear, where AC stands for air conduction (using sounds from a headphone on the ear, measuring overall hearing ability) and BC stands for bone conduction (the sound is given to the mastoid bone behind the ear, measuring the hearing ability of the inner ear - cochlea and auditory nerve). An example of an audiogram for one ear would be |65|65|35|40|45|55|0|10|25|40|50|. The first six values are AC thresholds (the faintest sound that the patient can hear in decibels) at 250, 500, 1000, 2000, 4000 and 8000 Hz, and the following five values are the BC thresholds for the same set of frequencies except 8000 Hz.
- Structured data, such as gender, date of birth, and hearing aid type, as in a typical relational database, e.g. |F|, |25-07-1991|, |ITENN|.
- Free text data / text comments, which are specific observations made about each patient e.g. |AT REV LT ITENL TO ITENN AS INSUFFICIENT GAIN-SOUNDED MUCH BETTER!|, which is shorthand for "At review, the left ITENL hearing aid was replaced by an ITENN hearing aid, as the old one had insufficient gain. The new one sounded much better."

## 3. K-Means Clustering

There are several potentially suitable clustering algorithms for audiology data, such as k-means clustering, hierarchical clustering, principal component analysis (PCA), or Self Organising Maps (SOM). We first consider k-means clustering, which is a non-probabilistic vector clustering method that uses iterative relocation to minimize within cluster variance. Here k is the number of cluster centroids. The k-means clustering algorithm can be described as follows (Bramer, 2007):

- Selection of the value of k
- Selection of any k objects and using them as the initial set of k centroids
- Assignment of each of the objects to the cluster with the closest to the centroid
- Recalculation of the centroids of the k clusters
- Repeating steps 3 and 4 until the centroids no longer move

### 3.1 Clustering of Audiograms by K-means

We performed clustering of hearing aid patient audiograms (right ear, AC, 250 to 8000 Hz) by using the k-means algorithm on 10,437, 1,316 and 13,136 records with text comments, diagnosis and hearing aid type respectively. We used all the records available in the database for each field under study for the experiments, keeping the criterion that none of the field values should be empty.

*Table 1.* Average Silhouette Values for clusters

| Clusters | Text comment | Diagnosis | Hearing aid type |
|---|---|---|---|
| 2 | 0.4005 | 0.5264 | 0.3934 |
| 3 | 0.3155 | 0.3679 | 0.2708 |
| 4 | 0.2760 | 0.3478 | 0.2618 |
| 5 | 0.2730 | 0.2749 | 0.2455 |
| 6 | 0.2524 | 0.2755 | 0.2381 |
| 22 | 0.1843 | 0.1793 | 0.1676 |

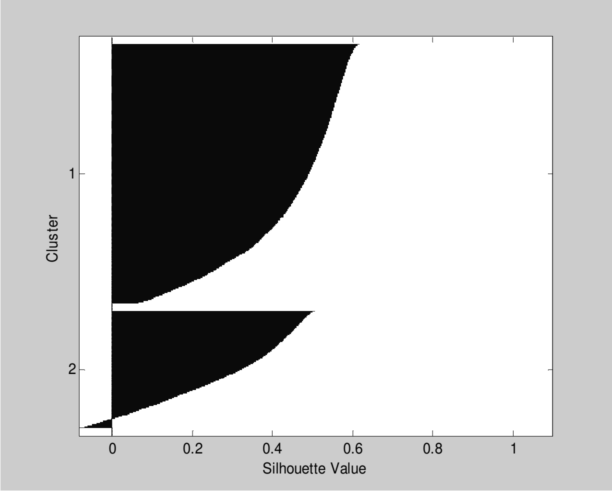


*Figure 1.* Silhouette plot with 2 clusters for hearing aid type of right ear air conduction frequencies.

To determine the correct number of clusters of audiograms we used average silhouette values. The silhouette plot displays a measure of closeness of points of a cluster with neighbouring clusters (Rao, & Kumar, 2009). The value ranges from +1 (indicating that points are very distant from neighbouring clusters) to -1 (indicating that points are assigned to the wrong cluster) while 0 indicates that points are not distinctly in one cluster or another (Rousseeuw, 1987; Matlab, 2010).

We calculated the average silhouette values for 2, 3, 4, 5, 6, and 22 clusters. The value of 22 was chosen by a rule of thumb which estimates the optimal number as clusters as the square root of half the number of records. The results are given in Table 1. Greater average silhouette values indicate that the clusters are better separated and average silhouette values do not always decrease with the number of clusters (Rousseeuw, 1987; Matlab, 2010). In our case, the optimal number of clusters based on right ear air conduction frequencies was 2 for each of three sets of records, as shown in Table 1. Figure 1 shows the silhouette plot obtained for 2 clusters for the set of records where hearing aid type was given. The other sets of records were those in which the text comment field was filled, and those in which the diagnosis field was filled in.

*Table 2.* Class Exemplars of each cluster of right ear air conduction frequencies for text comment

| | ac250 | ac500 | ac1K | ac2K | ac4K | ac8K |
|---|---|---|---|---|---|---|
| C1 | 73.66 | 73.00 | 74.99 | 80.48 | 91.08 | 108.21 |
| C2 | 35.17 | 33.56 | 35.87 | 43.26 | 55.90 | 66.50 |

*Table 3.* Class Exemplars of each cluster of right ear air conduction frequencies for diagnosis

| | ac250 | ac500 | ac1K | ac2K | ac4K | ac8K |
|---|---|---|---|---|---|---|
| C1 | 65.11 | 66.40 | 69.31 | 73.69 | 81.89 | 91.02 |
| C2 | 21.88 | 18.39 | 17.83 | 20.87 | 34.85 | 42.94 |

*Table 4.* Class Exemplars of each cluster of right ear air conduction frequencies for hearing aid type

| | ac250 | ac500 | ac1K | ac2K | ac4K | ac8K |
|---|---|---|---|---|---|---|
| C1 | 68.78 | 68.12 | 70.35 | 76.56 | 87.86 | 106.88 |
| C2 | 36.98 | 35.77 | 39.10 | 48.44 | 61.20 | 72.13 |

We calculated the class exemplar (cluster centroid) of each cluster, being the mean of the audiograms contained within each cluster, as shown in Tables 2 to 4. In Tables 2 and 4, the class exemplars show that cluster 1 consists of patients with severe hearing loss, and cluster 2 consists of patients with a mild to moderate hearing loss. In Table 3, cluster 1 corresponds to moderate to severe hearing loss and cluster 2 corresponds to normal or near-normal hearing. This is because in the majority of cases where a diagnosis was given, the diagnosis was tinnitus, which can occur even when there is little or no hearing loss.

### 3.2 Automatic Labeling of Clusters

Labeling clusters with the chi-squared technique is a reliable method (Manning, Raghavan, & Schütze, 2009). But, automatic cluster labeling has not been given much importance in the field of data mining (Tzerpos, 2001). Automatically labeling clusters gives the advantages of time reduction, better understanding and defining the purpose of each cluster (Maqbool, & Babri, 2006). After clustering the right ear air conduction audiograms, we used the chi-squared test to find which of the text keywords in the text comment, diagnosis and hearing aid type fields of the database were most and least typical of each cluster.

The Chi-squared test is a statistical non-parametric test which reveals associations between pairs of variables (fields of tables). It allows a comparison of frequencies found experimentally with those based on a theoretical model (Lucy, 2005; Oakes, & Farrow, 2007). It is calculated by determining the difference between a set of observed and expected frequencies within a population, and is given by the formula (Altman, 1991):

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where r is the number of unique terms in a particular field of the patient records (corresponding to rows in Table 6), and c is the number of clusters in the data found by the average silhouette values (Matlab, 2010), corresponding to columns in Table 6. The Chi-squared test is a simple test but is widely used in the medical domain. For example, it was successfully used in pharmacology by Oakes et al. (2001) to classify texts according to subtopics. We then produced a table for each field showing how often each of these words was associated with each cluster.

*Table 5.* Observed and Expected frequencies for right ear diagnosis

| Diagnosis | Cluster 1 | Cluster 2 | Row Total |
|---|---|---|---|
| TINNITUS | 171<br>(241.93)<br>[5030.54] | 954<br>(883.07)<br>[5030.54] | 1125 |
| OTHERS | 112<br>(41.07)<br>[5030.54] | 79<br>(149.93)<br>[5030.54] | 191 |
| Column Total | 283 | 1033 | 1316 |

Table 5 is the table produced for diagnoses occurring in the diagnosis field. Observed frequencies appear at the top of each cell, Expected frequencies are in ( ), and (Observed frequency – Expected frequency)$^2$ values are shown in [ ]. For example, if 171 of the diagnosis fields of the records of patients in cluster 1 contained the term 'TINNITUS', we would record a value of 171 for that term being associated with that cluster. These values were the "observed" values, denoted $O_{ij}$ in the formula above. The corresponding "expected" values $E_{ij}$ were found by the formula:

Row total x Column total / Grand Total

The row total for 'TINNITUS' diagnosis is the total number of times 'TINNITUS' hearing aids were prescribed to patients in two clusters = 171 + 954 = 1125. The column total for cluster 1 is the total number of patients assigned to cluster 1 over all diagnosis types = 283. The grand total is the total number of patient records in the study = 1316. Thus the "expected" number of 'TINNITUS' diagnosis in cluster 1 was 1125 * 283 / 1316 = 241.93. The significance of this is that the observed value is less than the expected value, suggesting that there is a negative degree of association between the 'TINNITUS' diagnosis and the severe hearing loss cluster. The remainder of the test is then performed to discover if this association is statistically significant.

Next the $O_{ij}$ and $E_{ij}$ values were used to calculate an overall chi-squared value for the relationship between keywords in the diagnosis field and cluster, using the formula above in Table 6. All the above steps were performed separately for the words in the text comments, diagnosis and hearing aid type fields, as shown in Table 6. From this data we could show, with 99.9% confidence, that these keywords were not randomly distributed, and that some keywords definitely are more associated with some clusters.

*Table 6.* Overall $\chi 2$

| Fields | Overall $\chi 2$ | Degrees of freedom (df) | P |
|---|---|---|---|
| Comments text | 4243.87 | 668 | P < 0.001 |
| Diagnosis | 182.52 | 1 | P < 0.001 |
| Hearing aid type | 5710.58 | 38 | P < 0.001 |

Having shown that overall, some keywords are more associated with some clusters, the next step was to discover exactly which individual keywords were most (and least) associated with each cluster. To do this, we considered the individual contributions of each word in each cluster to the overall chi-squared value for each text field, found by the formula

$$X^2 = \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

for each word in each cluster.

As the chi-squared test is unreliable for expected values of less than 5, for the diagnosis fields all words with such low expected values were grouped into a single class called 'OTHERS'.

Since we were in effect performing many individual statistical tests, it was necessary to use the Bonferroni correction (Altman, 1991) to control the rate of Type I errors where a word spuriously appears to be typical of a cluster. Since we wished to be 95% confident that a particular keyword was typical of a particular cluster, the corresponding significance level of 0.05 had to be divided by the number of simultaneous tests, i.e. the number of unique words times the number of clusters. In the case of words in the diagnosis fields, this gave a corrected significance level of 0.05 / (2 * 2) = 0.0125. Using West's chi-squared calculator (Chi-square calculator, 2010), for one degree of freedom we obtained a corresponding chi-squared threshold of 6.239. Thus we took all words in each cluster with individual contributions to the overall chi-squared value of over 6.239 to be significant at the 95% confidence level. The corresponding chi-squared thresholds were 11.65 for hearing aid type and 17 for the text comments.

Words associated with clusters with 95% confidence were deemed typical of those clusters if O > E, otherwise they were deemed atypical of those clusters. The words most typical and atypical of each cluster are shown in Tables 7 to 9. These automatically discovered words provided a suitable set of both positive and negative labels for each of the clusters. The labels seem intuitively reasonable. For example, in Table 7, it appears that the patients in cluster 2, the mild hearing loss group, were more concerned about tinnitus (ringing in the ears) than hearing loss. Thus the words tinnitus and masker (a machine for producing white noise to drown out tinnitus) were typical of this cluster and also are atypical of cluster 1, the severe hearing loss group. The hearing aid types associated with cluster 1 were those with high gain, while less powerful hearing aid types were negatively associated with this cluster. Similarly, in (Table 7) cluster 1, the atypical words "canc" (cancelled) and "dna" (did not attend) show that patients with severe hearing loss were less likely to cancel (or simply fail to attend) their appointments. 'Tinnitus' appears as 'tinnitu' and 'Suitable' appears as 'suitabl' in Table 7, since all the text was passed through Porter's (1980) stemmer for the removal of grammatical endings.

*Table 7.* Clusters for the records with text fields with positive and negative keywords.

| | Positive keywords | Negative keywords |
|---|---|---|
| C1 | audio, mould, be34, be52, be36, unmask, be54, sil, ref, tsa, gp, ca, OTHERS, rt, suitabl, be201 | masker, rev, tinnitu, appt, fta, help, review, aid, further, nfa, progress, 2000, ok, canc, counsel, cope, 2001, dna |
| C2 | masker, rev, tinnitu, appt, fta, help, review, aid, further, nfa | audio, mould, be34, be52, be36, unmask, be54, sil, ref |

*Table 8.* Clusters for the records with diagnosis fields with positive and negative keywords.

| | Positive keywords | Negative keywords |
|---|---|---|
| C1 | OTHERS | TINNITUS |
| C2 | None | OTHERS |

*Table 9.* Clusters for the records with hearing aid fields with positive and negative keywords.

| | Positive keywords | Negative keywords |
|---|---|---|
| C1 | BE34, BE52, BE36, BE54, ITEPN, BE201, PPCL, ITENN, PPC2, BE53, BE38, BE51, ITEPH2, BW83, BE35, PPC2D, PFPPCL, OTHERS, BE37 | ITEHN, ITEHH, BE19, ITENH, BE18, ITENL |
| C2 | ITEHN, ITEHH, BE19, ITENH, BE18 | BE34, BE52, BE36, BE54, ITEPN, BE201, PPCL, ITENN, PPC2, BE53, BE38, BE51, ITEPH2, BW83, BE35, PPC2D, PFPPCL, OTHERS |

## 4. Clustering of Audiograms by SOM

SOM (Self Organizing Map, also known as a Kohonen feature map) is an alternative multivariate neural technique which can examine interactions between a number of variables. SOM is an unsupervised learning process which clusters high dimensional data and gives output in groups or clusters. SOMs visualize or project high-dimensional data to low dimensions. They are used in pattern recognition, biological modeling, data compression and data mining. For example, Zehraoui and Bennani (2004) presented a SOM for sequence clustering and classification. Oakes et al. (2005) used SOM to cluster and classify unstructured and structured portions of audiology records.

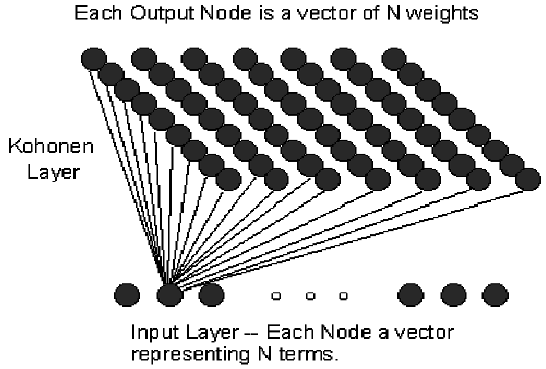

Each Output Node is a vector of N weights
Kohonen Layer
Input Layer -- Each Node a vector representing N terms.

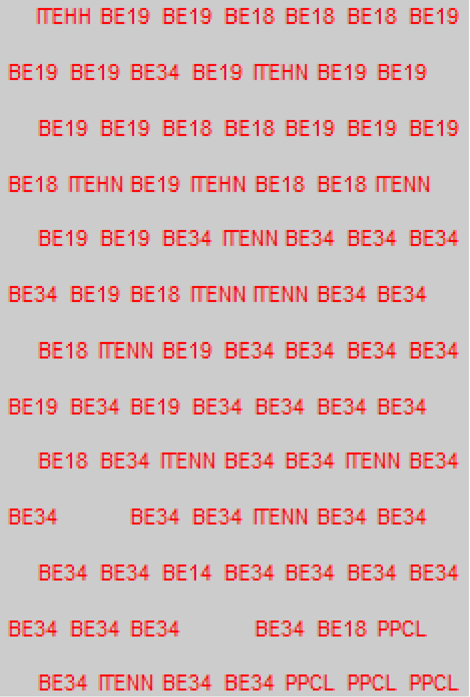
*Figure 2.* Kohonen SOM topology (Roussinov, & Chen, 1998)

Figure 2 shows a typical SOM architecture, in which all neurons are arranged on a fixed grid of the output layer and contain a weight vector (not shown in the figure) similar to the input dimensions. After training, each neuron represents different types of input data. Topological order is maintained in SOMs, this means the neurons that have similar weight in the input dimension are also close to each other in the SOM output map.

We used SOM for the clustering of our audiology records because of their strength in unsupervised learning and easy usage. In each case the input vectors, the basis of the clustering, were the patient audiograms. It should be noted that the values of quantization error and topographic error decrease as map size is increased. In Figure 3, the clustered labels of hearing aid type are shown and they can be seen to form clusters, for example, BE34 on the bottom left, BE19 on the top left, and PPCL in the bottom right corner.

## 5. Comparison between K-Means and SOM

SOM and k-means are complimentary techniques in that SOM gives a visual way of looking at individual patients while k-means finds the prototypical members of clusters of similar patients such as those with similar audiograms. In SOM we labeled each patient record by a single word such as hearing aid type, while the clusters produced by k-means are labeled using the chi-squared technique with number of keywords. However, exactly the same techniques could be used in both.



```
   ITEHH BE19 BE19 BE18 BE18 BE18 BE19
BE19 BE19 BE34 BE19 ITEHN BE19 BE19
   BE19 BE19 BE18 BE18 BE19 BE19 BE19
BE18 ITEHN BE19 ITEHN BE18 BE18 ITENN
   BE19 BE19 BE34 ITENN BE34 BE34 BE34
BE34 BE19 BE18 ITENN ITENN BE34 BE34
   BE18 ITENN BE19 BE34 BE34 BE34 BE34
BE19 BE34 BE19 BE34 BE34 BE34 BE34
   BE18 BE34 ITENN BE34 BE34 ITENN BE34
BE34      BE34 BE34 ITENN BE34 BE34
   BE34 BE34 BE14 BE34 BE34 BE34 BE34
BE34 BE34 BE34      BE34 BE18 PPCL
   BE34 ITENN BE34 BE34 PPCL PPCL PPCL
```

*Figure 3.* A snapshot of clustering of right ear hearing aids from right ear air conduction frequencies.

The findings of k-means with automatic cluster labeling and SOM confirmed each other. Looking at the outcomes of SOM and k-means in Table 9 and Figure 3, we note that the hearing aid types BE34 and ITENN are both positive keywords of cluster 1 by k-means and they are also found close together in the SOM output. This is reasonable, since both these hearing aids have similar acoustic gain. Similarly, the keywords ITEHN, BE19, BE18 can be seen both in k-means cluster 2 and also in adjacent positions in the SOM.

## 6. Conclusion

We have clustered audiology patient records and assigned text labels automatically to help in the interpretation of clusters. These text labels will be helpful in the construction of an audiology decision support system for the selection of hearing aids. In one experiment we clustered the audiograms by k-means clustering, and then used the chi-squared test to assign labels taken from the text fields in the database. In our second experiment, we used SOMs to cluster audiograms individually labeled with the type of hearing aid selected for each patient. In the experiments reported here, we have used AC thresholds alone as the basis for clustering. In future we will use a combination of both AC and BC, since the difference between the two yields information about the cause of deafness. We will also produce automatic labels for clusters of audiograms produced by PCA and hierarchical clustering, using the method of Maqbool and Babri (Maqbool, & Babri, 2006). Although our data set consists only of audiology records, these are somewhat representative of medical records in general, as they consist of numeric, structured text and unstructured textual fields.

## Acknowledgment

We wish to thank Maurice Hawthorne, 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

Altman, D. G. (1991). *Practical Statistics for Medical Research*, Chapman & Hall, 241-248.

Bramer, M. (2007). *Principles of Data Mining*. Springer, 224-231.

Hung, C. (2004). *An adaptive SOM model for document clustering using hybrid neural techniques*, Thesis PhD – University of Sunderland, 32-33.

Ke, S-W., Oakes, M., & Bowerman, C. (2005). Mining personal data collections to discover categories and category labels, *RANLP Text Mining Workshop*, Borovets, Bulgaria, 17-22.

Lucy, D. (2005). *Introduction to statistics for forensic scientists*. John Wiley & Sons Ltd, 45-52.

Manning, C. D., Raghavan, P., & Schütze, H. (2009). *Introduction to information retrieval*, Cambridge University Press, Cambridge, England, 271-278.

Maqbool, O., & Babri, H. A. (2006). Automated software clustering: An insight using cluster labels, *The Journal of Systems Software*. Elsevier Inc., 1632-1648.

Matlab: Statistical toolbox, K-means clustering. Retrieved March 23, 2010, from http://www.mathworks.com/access/helpdesk/help/toolbox/stats/bq_679x-18.html

Oakes, M., Cox, S., & Wermter, S. (2005). Data mining audiology records with the Chi-squared test and self-organising maps, *22nd British National Conference on Databases*, D. Nelson et al., (Eds.), University of Sunderland Press, 123-130.

Oakes, M., Gaizauskas, R., Fowkes, H. et al., (2001). Comparison between a method based on the chi-square test and a support vector machine for document classification, *Proceedings of ACMSIGIR*, New Orleans, 440-441.

Oakes, M. P., & Farrow, M. (2007). Use of the chi-squared test to examine vocabulary differences in English language corpora representing seven different countries, *Literary & Linguistic Computing,* 22(1) 85- 99.

Porter, M. F. (1980). *An algorithm for suffix stripping, Program,* Vol. 14, No. 3, 130-137.

Principe, J. C., Euliano, N. R., & Lefebvre, W. C. (2000). *Neural and Adaptive Systems: Fundamentals through Simulations*, John Wiley & Sons, INC., 341-344.

Rao, V. S. H. & Kumar, M. N. (2009). Estimation of the parameters of an infectious disease model using neural networks, *Nonlinear Analysis: Real World Applications*, Elsevier Ltd., (Article in press).

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics 20*, 53-65.

Roussinov, D. G., & Chen, H. (1998). A scalable self-organizing map algorithm for textual classification: a neural network approach to automatic thesaurus generation, *Communication and Cognition in Artificial Intelligence Journal (CC-AI), Volume 15, Number 1-2*, 81-111.

Tzerpos, V. (2001). *Comprehension-driven software clustering*, Thesis PhD – University of Toronto.

Wermter, S., & Sun, R. (2000). Hybrid Neural Systems. *Lecture Notes in Artifical Intelligence 1778*, Springer.

West's Chi-square calculator. Retrieved January 22, 2010, from http://www.stat.tamu.edu/~west/applets/chisqdemo.html

Zahraoui, F., & Bennani, Y. (2004). M-SOM-ART: growing self organizing map for sequence clustering and classification, *16th European conference on AI*, Valencia, Spain, 564-570.