From Mining Tinnitus Database to Tinnitus Decision-Support System, Initial Study

Pamela Thompson, Xin Zhang, Wenxin Jiang, Zbigniew W. Ras
Department of Computer Science, University of North Carolina at Charlotte
pthompso@catawba.edu, {xinzhang, wjiang3, ras}@uncc.edu

Abstract

Many definitions for tinnitus exist and causes and treatments are plentiful, yet not completely understood. A database of eleven tables, as many as 555 unique patien tuples and numerous time-stamped and other features was obtained for knowledge discovery related to causes and treatments of tinnitus. The paper describes the knowledge discovery and machine learning process and introduces several new temporal features to improve tinnitus evaluation, outcomes analysis, and overall understanding. Through automated analysis it is the goal of the authors to determine unknown yet potentially useful attributes related to tinnitus research.

1. Introduction

A tinnitus patient database of eleven tables and 555 patient tuples was obtained from Dr. Pawel J. Jastreboff, a leading tinnitus researcher at Emory University in Atlanta, GA. The authors have a research goal of performing data mining on the tinnitus database in order to determine rules, associations and create new and useful features to assist with understanding and validation of diagnosis and treatment outcomes. In order to perform data mining, it was first necessary to gain knowledge regarding the domain of the data.

Tinnitus, sometimes called "ringing in the ears", affects a significant portion of the population. Some estimates show the portion of the population in the United States affected by tinnitus to be 40 million, with approximately 10 million of these considering their problem significant [1]. Many definitions exist for tinnitus. One definition of tinnitus relevant to this research is ". . . the perception of sound that results exclusively from activity within the nervous system without any corresponding mechanical, vibratory activity within the cochlea, and not related to external stimulation of any kind" [2]. Hyperacusis or decreased sound tolerance frequently accompanies tinnitus and can include symptoms of misophonia (strong dislike of sound) or phonophobia (fear of sound). Physiological causes of tinnitus can be difficult or impossible to determine, and treatment approaches vary.

Tinnitus Retraining Therapy (TRT), developed by Dr. Jastreboff, is one treatment model with a high rate of success and is based on a neurophysical approach to treatment. TRT "cures" tinnitus by building on its association with many centers throughout the nervous system including the limbic and autonomic systems. The limbic nervous system (emotions) controls fear, thirst, hunger, joy and happiness. It is connected with all sensory systems. The autonomic nervous system controls such functions as breathing, heart rate and hormones. When the limbic system becomes involved with tinnitus, symptoms may worsen and affect the autonomic nervous system [3]. TRT combines counseling and sound habituation to successfully treat a majority of patients.

Conceptually, habituation refers to a decreased response to the tinnitus stimulus due to exposure to a different stimulus [4]. Degree of habituation determines treatment success, yet greater understanding of why this success occurs and validation of the TRT technique will be useful [5].

The treatment requires a preliminary medical examination, completion of an Initial Interview Questionnaire for patient categorization, audiological testing, a visit questionnaire referred to as a Tinnitus Handicap Inventory, tracking of instruments, and a follow-up questionnaire. Data from a sample of 555 patients was originally presented in a relational database consisting of eleven tables. Patient tuples were related to one to many tuples in other tables based on patient visits through the course of treatment. Tuples with data related to treatments during visits were uniquely identified by patient id and visit number and date, enabling temporal treatment of data. The authors focused on cleansing and analysis of existing data, along with automating the discovery of new and useful features in order to improve classification and understanding of tinnitus diagnosis and improvement.

2. Treatment process

The original database represented many challenges in successful mining and analysis. The database was created from manual forms from patient visits and treatments transcribed to a relational database format over a period of years. In order to perform research, it was important to understand the domain knowledge related to the problem from the areas of otology, psychology, and computer science. Based on the domain knowledge, a determination had to be made on useful features to the problem, then data had to be consolidated and cleansed with many discrepancies resolved programmatically. Some similar data was stored in different formats, and contained inconsistencies. Null values were programmatically removed and generally not included in the aggregate table.

An additional problem was presented with the task of creating a single table of data for mining purposes. Relationships among multiple tables were based on a patient id that was represented in different formats, and a visit date and number related to many of the tuples was not consistent across tables.

3. Feature extraction

Useful features pertinent to data mining were extracted and transformed in preparation for analysis. The goal was to extract those features that described the situation of the patient based on the behavior of the attributes over time, and to transform, discretize and classify them in new ways where useful. Many algorithms exist for discretization, yet in this research expert domain knowledge provided the basis for many of the discretization algorithms. This section will identify the resulting features along with a description of the transformation performed.

The patient id was standardized across tables. Visit number and visit data were transformed to total visits (representing number of visits) and length of visit, a continuous temporal feature that determined time span in number of days between first and last visits.

Patient data related to visits includes a determination of the problem in order of importance, stored as various combinations of the letters "T" for tinnitus, "H" for Hyperacusis, and L for "Loudness Discomfort". Only the first and last of these (First P and last P) are stored in two separate attributes in the analysis table related to first and last visit.

Patient category represents the classification of the patient and is represented twice: first by original category and second by category of treatment. This feature is represented by a range of scores from 0 to 4 where 0 represents tinnitus as a minimal problem, 1 represents tinnitus as a significant problem, 2 represents tinnitus as a significant problem and hearing loss a significant subjective problem, 3 represents tinnitus is irrelevant and hyperacusis is a significant problem with hearing difficulties irrevalent, and 4 with tinnitus represents prolonged tinnitus with hearing difficulties irrelevant, and 4 with hyperacusis represents tinnitus and hearing difficulties irrelevant with prolonged exacerbation of hyperacusis caused by sound [6]. Two patient categories are stored in the final table (C and Cc), the first category assigned representing the diagnosis and the last category assigned representing the final determination of the patient problem. Assigning the patient to a category is important to treatment success, and a successfully treated patient will move toward category 0 [7]. Some analysis was performed based on this feature.

The Tinnitus Handicap Inventory score (T Score) was discretized based on domain knowedge. Lower T Scores are better. The difference in T Score from first to last visit was calculated and discretized from "a" to "e" with "a" for good to "e" for bad representing the improvement in the patient. This feature is stored as Category of T score.

The standard deviation of audiological testing features related to loudness discomfort levels was derived and stored in various attributes in the analysis table. Loudness discomfort level is a measure of decreased sound tolerance as indicated by hyperacusis or discomfort to sound, misophonia or dislike of sound and phonophobia or fear of sound. Loudness discomfort levels change with treatment and patient improvement, unlike other audiological features. For this reason the audiological data related to loudness discomfort levels is included in analysis.

Finally, information on instruments and models of equipment used by the patient is stored in text format in the analysis table.

Types of learning desired include classification learning to help with classifying unseen examples, association learning to determine any association among features (largely statistical) and clustering to seek groups of examples that belong together.

Temporal features have been widely used in various research areas, such as stream tracer study [8], music sound classification [9], and business intelligence [10]. It is especially important in the light of the tinnitus treatment process. Evolution of sound loudness discomfort level parameters in time is essential for treatments, therefore it should be reflected in treatment features as well. The discovered temporal patterns may better express treatment process than static features, especially considering that the standard deviation and mean value of the sound loudness discomfort level features can be very similar for sounds representing the same type of Tinnitus treatment category, whereas changeability of sound features with tolerance levels for the same type of patients makes recovery of one type of patients dissimilar. New temporal features include:

Sound level Centroid:

$$C = \frac{\sum_{n=1}^{length(T)} n / length(T) \cdot V(n)}{\sum_{n=1}^{length(T)} V(n)}$$

C is the gravity center of the sound level feature V, V(n) is the sound level feature V in the nth visit, T is the total visits.

Sound level Spread:

$$S = \sqrt{\frac{\sum_{n=1}^{length(T)} (n / length(T) - C)^2 \cdot V(n)}{\sum_{n=1}^{length(T)} V(n)}}$$

Recovery Rate:

$$\frac{V_0 - V_k}{T_k - T_0}, k \in \min\{V_i\}, i \in [0, N]$$

In Recovery Rate, V represents the total score from the Tinnitus Handicap Inventory in a patient visit. V_o is the first score recorded from the Inventory during the patient initial visit. V_k represents the minimum total score which is the best out of the vector of the scores across visits. V_o should be greater meaning the patient is worse based on the Inventory from the first visit. T_k is the date that has the minimum total score, T_o is the date that relates to Vo. A large recovery rate score can mean a greater improvement over a shorter period of time. XY scatter plots were constructed using recovery rate compared to patient category, and recovery rate compared to treatment category in order to determine interesting patterns in the data.

4. Classifier

Decision tree study was performed using C4.5, a system that incorporates improvements to the ID3 algorithm for decision tree induction. C4.5 includes improved methods for handling numeric attributes and missing values, and generates decision rules from the trees [11]. In this study C4.5 was used in order to evaluate the recovery rate by all of the attributes in the research database and to learn tinnitus by loudness discomfort levels, new temporal features and problem types.

5. Experiment and results

In this research, we performed two different experiments: Experiment#1 explored Tinnitus treatment records of 253 patients and applied 126 attributes to investigate the association between treatment factors and recovery; Experiment#2 explored 229 records and applied 16 attributes to investigate the nature of tinnitus with respect to hearing measurements. All classifiers were 10-fold cross validation with a split of 90% training and 10% testing. We used WEKA for all classifications.

Preliminary research results show several interesting rules resulting from decision tree analysis:

5.1. Experiment#1:

- (Category of treatment = C1) $\wedge$ (R50 >12.5) $\wedge$ (R3 <=15)==> improvement is neutral

The support of the rules is 10, the accuracy is 90.9%. It means that if treatment category chosen by patient is C1 then when R50 parameter is above 12.5 and average of R3 is less or equals to 15 then the recovery is neutral.

- (Category of treatment = C2) ==> good

The support of the rules is 44, the accuracy is 74.6%. It means that if category of treatment chosen by patient is C2 then Improvement is good.

- (Category of treatment = C3) $\wedge$ (Model = BTE)==>good

The support of the rules is 17, the accuracy is 100.0%.

5.2. Experiment#2:

- 40>Lr50>19 ==>Somehow has tinnitus all of the time

The support of the rules is 27, the accuracy is 100.0%. It means that if Lr50 is in range of 19 to 40, somehow the patient has tinnitus all the time, where the tinnitus may not be a major problem.

Scatter plot analysis shows when recovery rate is compared to patient and treatment category in XY scatter plot analysis, both patient and treatment category 4 shows a smaller rate of recovery value possibly indicating slower or reduced treatment success.

6. Conclusion

Data mining is a promising and useful technique for the discovery of new knowledge related to tinnitus causes and cures. Continued research should provide additional learning.

7. Future work

This research is in the early stages, and many possibilities exist for further analysis and learning. The data set can be expanded to include medical data related to patients. Textual features need to be classified and discretized in order to be used in discovery. For example, the cause of tinnitus according to the patient is a text field that can be classified and used in knowledge discovery. A text field listing medications can also be considered. A determination of features that can be tied to emotions will be useful in analysis and determination of sound therapy tied to emotions. In therapy, it is important that sounds used do not cause annoyance as the involvement of the emotional (limbic) system is influential in perception of tinnitus and its severity. Finding sounds that improve emotions can shorten the treatment time (as measured by the new temporal feature recovery rate) particularly if there is some way to prescribe certain sounds and monitor the nervous system as treatment continues.

8. Acknowledgment

We wish to acknowledge the significant contribution to this research from Dr. Pawel J. Jastreboff, PhD, ScD, Emory University School of Medicine.

9. References

[1] Folmer, Robert L (2002). Long-term Reductions in Tinnitus Severity. BMC Ear, Nose and Throat Disorders, p. 1.

[2] Jastreboff, Pawel J. and Jonathan W. P. Hazell (2004). Tinnitus Retraining Therapy: Implementing the Neurophysical Model, Cambridge University Press, p. 4.

[3] Jastreboff, Pawel J. and Jonathan W. P. Hazell (2004). Tinnitus Retraining Therapy: Implementing the Neurophysical Model, Cambridge University Press, p. 12-13.

[4] Henry, James A., Margaret M. Jastreboff, Pawel J. Jastreboff, Martin A. Schechter, and Stephen A. Fausti (2003). Guide to Conducting Tinnitus Retraining Therapy Initial and Follow-Up Interviews. Journal of Rehabilitation Research and Development, Vol. 40, No. 2, March/April 2003, p. 159-160.

[5] Baguley, David M. What Progress Have We Made With Tinnitus?, Acta Oto-Laryngologica, 2006; 126: 5.

[6] Henry, James A., Margaret M. Jastreboff, Pawel J. Jastreboff, Martin A. Schechter, and Stephen A. Fausti (2003). Guide to Conducting Tinnitus Retraining Therapy Initial and Follow-Up Interviews. Journal of Rehabilitation Research and Development, Vol. 40, No. 2, March/April 2003, p. 161.

[7] Jastreboff, Pawel J. and Jonathan W. P. Hazell (2004). Tinnitus Retraining Therapy: Implementing the Neurophysical Model, Cambridge University Press, p. 84.

[8] Waldon, Michael G. (2004) Estimation of Average Stream Velocity, in J. Hydr. Engrg., Volume 130, Issue 11, pp. 1119-1122 (Nov. 2004).

[9] Zhang, X., Ras, Z.W. (2006). Differentiated Harmonic Feature Analysis on Music Information Retrieval for Instrument Recognition, Proceedings of IEEE International Conference on Granular Computing (IEEE GrC 2006), May 10-12, Atlanta, Georgia, 578-581.

[10] Povinelli, Richard J., Xin Feng. (1998) "Temporal Pattern Identification of Time Series Data using Pattern Wavelets and Genetic Algorithms" Artificial Neural Networks in Engineering, Proceedings, 691-696.data using pattern wavelets and genetic algorithms

[11] Witten, Ian H. and Eibe Frank (2005). Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann Publishers, 2nd Edition, p. 105.