Mining Tinnitus Data based on Clustering and new Temporal Features

Xin Zhang, Pamela Thompson, Zbigniew W. Raś, Pawel Jastreboff

Abstract Tinnitus problems affect a significant portion of the population and are difficult to treat. Sound therapy for Tinnitus is a promising, expensive, and complex treatment, where the complete process may span from several months to a couple of years. The goal of this research is to explore different combinations of important factors leading to a significant recovery, and their relationships to different category of Tinnitus problems. Our findings are extracted from the data stored in a clinical database, where confidential information had been stripped off. The domain knowledge spans different disciplines such as otology as well as audiology. Complexities were encountered with temporal data and text data of certain features. New temporal features together with rule generating techniques and clustering methods are presented with a ultimate goal to explore the relationships among the treatment factors and to learn the essence of Tinnitus problems.

Xin Zhang
University of North Carolina at Pembroke, Dept. of Math. Comp. Science, Pembroke, NC 28372, USA e-mail: xin.zhang@uncp.edu

Pamela Thompson
University of North Carolina, Dept. of Computer Science, Charlotte, NC 28223, USA e-mail: pthompso@catawba.edu

Zbigniew W. Raś
University of North Carolina, Dept. of Computer Science, Charlotte, NC 28223, USA & Warsaw University of Technology, Institute of Comp. Science, 00-665 Warsaw, Poland e-mail: ras@uncc.edu

Pawel Jastreboff
Emory University School of Medicine, Dept. of Otolaryngology, Atlanta, GA 30322, USA e-mail: pjastre@emory.edu

1 Introduction

Tinnitus affects a significant portion of the population. It is rather a symptom than a disease. The definition of Tinnitus based on its individual characters was well discussed by Jastreboff [5]. For many years, Tinnitus was believed impossible to be treated. Not until recently, Tinnitus Retraining Therapy (TRT), which has been developed by Jastreboff (for details, see [5] and [4]) is a promising, complex, and expensive treatment based on the neurophysical model of tinnitus, and is aimed at inducing and sustaining habituation of tinnitus-evoked reactions and tinnitus perception (neurophysiology is a branch of science focusing on the physiological aspect of nervous system function (for details, see [5])). On one hand, TRT has provided relief for many patients as well as generated a high volume of medical data in the format of matrix-in-matrix, which is not suitable for traditional data mining algorithms. On the other hand, due to the fact that objective methods are lacking to detect Tinnitus symptoms, it is also interesting for the clinical doctors to be able to learn the essences of Tinnitus through the audiological evaluation data. Thus, the authors started their initial research by exploring the relationships among the complex factors of the treatment and recovery patterns in different categories of patients for the purpose of optimizing the treatment process as well as learning the essence of the Tinnitus problems.

The rest of this section will focus on the basic domain knowledge necessary to understand TRT and its data collection.

1.1 TRT Background

The domain knowledge for tinnitus involves many disciplines, primarily including otology and audiology. Tinnitus appears to be caused by a variety of factors including exposure to loud noises, head trauma, and a variety of diseases. An interesting fact is that Tinnitus can be induced in 94% of the population by a few minutes of sound deprivation [2]. Decreased sound tolerance frequently accompanies tinnitus and can include symptoms of hyperacusis (an abnormal enhancement of signal within the auditory pathways), misophonia (a strong dislike of sound) or phonophobia (a fear of sound) [5]. Past approaches to treatment tend to have been based on anecdotal observations and treatment often focused on tinnitus suppression. Currently a wide variety of approaches are utilized, ranging from sound use to drugs or electrical or magnetically stimulation of the auditory cortex. Jastreboff [4] proposed an important new model (hence treatment) for tinnitus that focuses on the phantom aspects of tinnitus with tinnitus resulting exclusively from activity within the nervous system that is not related to corresponding activity with the cochlea or external stimulation. The model furthermore stresses that in cases of clinically-significant tinnitus, various structures in the brain, particularly the limbic and autonomic nervous system, prefrontal cortex, and reticular formations play a dominant role with the auditory system being secondary. Tinnitus Retraining Therapy (TRT), developed by Jastreboff, is a treatment model with a high rate of success (over 80% of the cases) and is based on the neurophysical model of tinnitus. Tinnitus Retraining Therapy "cures" tinnitus-evoked reactions by retraining its association with specific centers throughout the nervous system, particularly the limbic and autonomic systems. The limbic nervous system (emotions) controls fear, thirst, hunger, joy and happiness and is involved in learning, memory, and stress. The limbic nervous system is connected with all sensory systems. The autonomic nervous system controls functions of the brain and the body over which we have limited control, e.g., heart beating, blood pressure, and release of hormones. The limbic and autonomic nervous systems are involved in stress, annoyance, anxiety etc. When these systems become activated by tinnitus-related neuronal activity (tinnitus signal) negative symptoms are evoked [5]. Unfortunately, many patients seeking treatment other than TRT are often told that nothing can be done about their tinnitus. This can have the negative effect of enhancing the limbic nervous system reactions, which then can cause strengthening of the negative effect of the tinnitus on a patient (see Fig. 1: [5]).

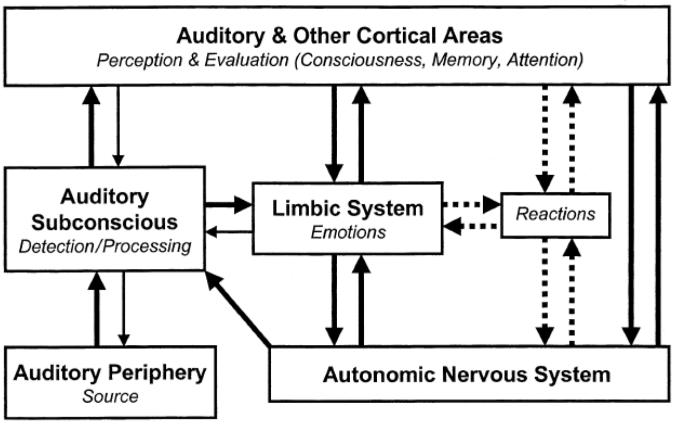

Fig. 1 Block diagram of the neurophysiological model of tinnitus

TRT is aimed at evoking and sustaining habituation of tinnitus-evoked reactions and tinnitus perception. Degree of habituation determines treatment success, yet greater understanding of why this success occurs and validation of the TRT technique will be useful [1]. The ultimate goal is to lessen or eliminate the impact of tinnitus on the patient's life [5]. Jastreboff has observed statistically significant improvement after three months and treatment lasts approximately 12-18 months.

1.2 TRT Data Collection

Tinnitus Retraining Therapy combines medical evaluation, counseling and sound therapy to successfully treat a majority of patients. Based on a questionnaire from the patient as well as an audiological test, a preliminary medical evaluation of patients is required before beginning TRT. Sensitive data from the medical evaluation is not contained in the tinnitus database O presented to the authors (except that Jastreboff maintains this data). Much of this data would contain information subject to privacy concerns, a consideration of all researchers engaged in medical database exploration. Some medical information, however, is included in the tinnitus database such as a list of medications the patient may take and other conditions that might be present, such as diabetes.

Category	Impact of Life	Tinnitus	Subjective Hearing Loss	Hyperacusis	Prolonged Sound-Induced Exacerbation	Treatment
0	Low	Present	-	-	-	Abbreviated counseling
1	High	Present	-	-	-	Sound generators set at mixing point
2	High	Present	Present	-	-	Hearing aid with stress on enrichment of the auditory background
3	High	Not relevant	Not relevant	Present	-	Sound generators set above threshold of hearing
4	High	Not relevant	Not relevant	Present	Present	Sound generators set at the threshold; very slow increase of sound level

Fig. 2 Patient Categories

Patient categorization is performed after the completion of the medical evaluation, the structured interview guided by special forms and audiological evaluation, this interview collects data on many aspects of the patient's tinnitus, sound tolerance, and hearing loss. The interview also helps determine the relative contribution of hyperacusis and misophonia. A set of questions relate to activities prevented or affected (concentration, sleep, work, etc.) for tinnitus and sound tolerance, levels of severity, annoyance, effect on life, and many others. All responses are included in the database. As a part of audiological testing left and right ear pitch, loudness discomfort levels, and suppressibility is determined. Based on all gathered information a patient category is assigned (see Fig. 2). A patient's overall symptom degree is evaluated based on the summation of each individual symptom level, where a higher value means a worse situation. During a medical treatment, comments may be recorded by the doctors or nurses together with a treatment category. Beside surveys for the purpose of symptom evaluation, the TRT process also includes sound therapy, which involves sound parameters for the ear(s), such as mixing point, suppression mask, stochastic resonance (adding low level signal noises), and various sound levels that are applied to a patient. The TRT emphasizes on working on the principle of differences of the stimuli from the background based on the fact that the perceived strength of a signal has no direct association with the physical strength of a stimulus, using a functional dependence of habituation effectiveness model, in Fig.

3. Therefore, once a partial reversal of hyperacusis is achieved, the sound level can be increased rapidly to address tinnitus directly.

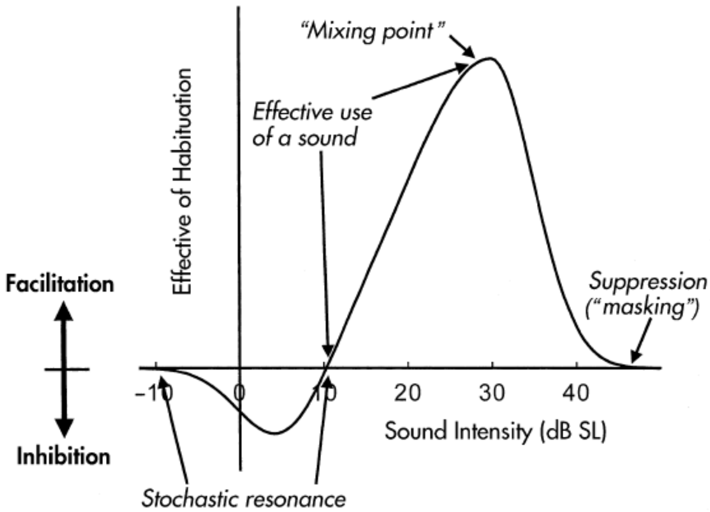

Fig. 3 Functional dependence of habituation effectiveness on physical intensity of a sound

2 Information Retrieval

This paper involves the construction of a database D in the desirable format for classifiers construction based on the database O that was mentioned in the previous section. It uses the term "attribute" to refer to a column in the table from the database O and the term "feature" to refer to a column in the database D. Also, due to the intuition of the process in each visit, recovery of any patient with only one visit cannot be evaluated. Therefore, such records had been removed during the experiments. Features have been developed in this research for the following types of data: text, continuous temporal data, categorical temporal data, among which continuous temporal data have standard statistical features (such as average as well as standard deviation) and new temporal features, as shown in Fig. 4.

Based on subject type, the medical record information for TRT can be grouped into two types: one is to describe properties of a patient, which are relatively stable (e.g., gender, age, occupation, marriage status, etc.); another one is to store properties of a visit of a patient, which may change relatively more frequently than the former type (e.g., total recovery score, drug prescript, clinical doctor's comments, audiological parameters of sound therapy, etc.).

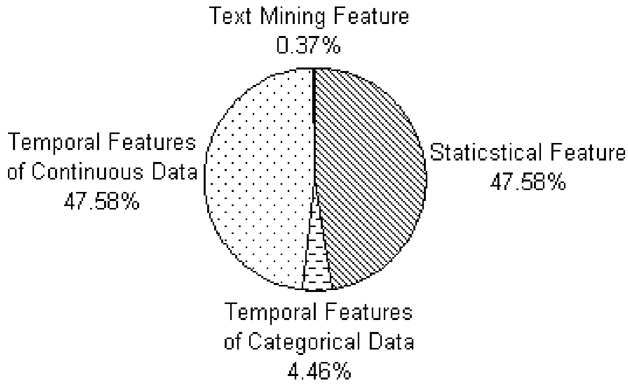

Fig. 4 Features in the Database O

In the light that total visit of each patient during the TRT process varies from several to, sometimes, over fifteen, the authors of this paper developed different formulas and algorithms to capture such subtle changes about visit over time and to transform the sparse data into an ideal data format for traditional classifiers, where each observation describes a patient.

The rest of this section will introduce the features for those three different data types respectively.

2.1 Mining Text Data

Many of the attributes in the original database that are stored in text format may contain important information which may have correlation to the new feature tinnitus recovery rate, and to the overall evaluation of treatment method. Some preliminary work has been done in this area with several text fields that indicate the cause or origination of tinnitus.

Text mining (also referred to as text classification) involves identifying the relationship between business categories and the text data (words and phrases). For details, see [14]. This allows the discovery of key terms in text data and facilitates automatic identification of text that is ”interesting”. The authors designed a new Boolean feature that indicates if tinnitus was induced by exposure to a loud noise. When performing extraction on data entered in text format it is important to recognize that the data may have been entered in a non-systematic manner and may have been entered by multiple individuals. The authors are making the assumption that careful text mining is worthwhile on the tinnitus database in many of the comment style attributes. The following are the text mining steps that were used:

- Term extraction transformation was used which performs such tasks as tokenizing text, tagging words, stemming words, and normalizing words. By this trans-

formation, 60 frequent terms were determined from a text attribute that describes how tinnitus was induced.
- After reviewing these terms, some terms were determined to be inconsequential in the domain. These terms were classified as noise words as they occurred with high frequency. These terms were then added to the exclusion terms list, which is used as a reference table for the second run of the term extraction transformation.
- The second Term extraction transformation resulted in 14 terms which are related to the Tinnitus induced reason of ”noise exposure”. The terms included in the dictionary table are ”concert”, ”noise”, ”acoustic”, ”exposure”, ”music”, ”loud”, ”gun”, ”explosion”, ”pistol”, ”band”, ”headphone”, ”firecracker”, ”fireworks”, and ”military”.
- Fuzzy-lookup transformation was applied which uses fuzzy matching to return close matches from the dictionary table to extract keywords/phrases into the attribute ”keywords”. This attribute indicates whether the induced reason for Tinnitus is related to exposure to a noise of some type. As mentioned previously, it is well recognized that noise in general and impulse noise in particular are the most common factors evoking tinnitus.

After adding this new attribute to the table, decision tree algorithms were applied in order to produce relevant rules. Twenty-nine patients have the value of true for the attribute ”keywords”, where each observation was associated to a patient identification number. The patients identified by this process can be considered clearly to have tinnitus induced by noise exposure. The main purpose of the text mining process is to add extra information to the study in order to improve the accuracy of the prediction model. While expanding the dictionary table will help improve the knowledge gained, relaxation of the rules used for text extraction may serve to enhance the rate of positive responses. This will be addressed in future work, along with more text attributes such as patient occupation, prescription medications, and others.

2.2 Temporal Feature Design for Continuous Data

TRT is a complex process with bunches of treatment as well as symptom parameters including not only drug prescriptions, but also audiological therapy, counseling, and evaluations, during which all para-meter values may be subjective to change over time for each individual patient. Therefore, the treatment-record database is sort of matrix-in-matrix format, which is not suitable for traditional data mining algorithms.

Temporal features have been widely used to describe subtle changes of continuous data over time in various research areas, such as stream tracer study [9], music sound classification [11], and business intelligence [6]. It is especially important in the light of the tinnitus treatment process because examining the relationship of patient categorization, the total scores, and audiological test results over time may be beneficial to gaining new understanding of the treatment process. Evolution of sound loudness discomfort level parameters in time is essential for evaluation of treatment outcome for decreased sound tolerance, but irrelevant for tinnitus; 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 is a feature of center of gravity of an audiological therapy parameter over a sequence of visits. It is defined as a visit-weighted centroid.

$$\left\{ V \in \varphi \mid C = \frac{\sum_{n=1}^{length(T)} n / length(T) \cdot V(n)}{\sum_{n=1}^{length(T)} V(n)}) \right\}$$

where φ represents the group of audiological features in the therapy, C is the gravity center of the sound audiological feature V, $V(n)$ is an audiological therapy feature V in the nth visit, T means the total number of visits.

Sound level Spread S is a feature of the Root of Mean Squared deviation of an audiological therapy feature over a sequence of visits with respect to its center of gravity.

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

where φ represents the group of audiological features in the therapy, S is the spread of the audiological feature V, $V(n)$ is a sound level feature V in the nth visit, T means the total number of visits.

Recovery Rate R describes the recovery over time.

$$\left\{ V \in \varphi \mid R = \frac{V_0 - V_k}{T_k - T_0}, k \in \min\{V_i\}, i \in [0, N] \right\}$$

where V represents the total score from the Tinnitus Handicap Inventory in a patient visit. V_0 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_0 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_0 is the date that relates to V_0.

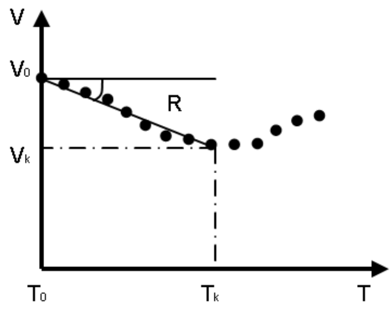

Fig. 5 Recovery Rate

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 against patient category, and recovery rate compared to against treatment category in order to explore potentially interesting patterns and relationships among those dimensions.

2.3 Temporal Feature Design or Categorical Data

During a period of medical treatment, a doctor may change the treatment from one category to another based on the specifics of recovery of the patients; the symptoms of a patient may vary as a result of the treatment; more so, the category of patient may change over time (e.g., hyperacusis can be totally eliminated and consequently the patient may move from category 3 to 1). Other typical categorical features, which may change over time in the database O include sound-instrument types as well as visiting frequencies. Statistical and econometric approaches to describe categorical data have been well discussed by Powers [7]. In our database, statistical features such as the most frequent pattern, the first pattern and the last pattern were used to describe the changes of categorical data over time.

Most frequent pattern MFP counts the pattern, which occurs most frequently for a particular patient. First and last pattern FP/LP represents the initial and final state of a categorical attribute respectively.

2.4 System Overview

Our TRT data mining system consists of five parts: data cleaning, feature extraction, K-means clustering, classification tree construction, and rules generation, as shown in Fig. 6.

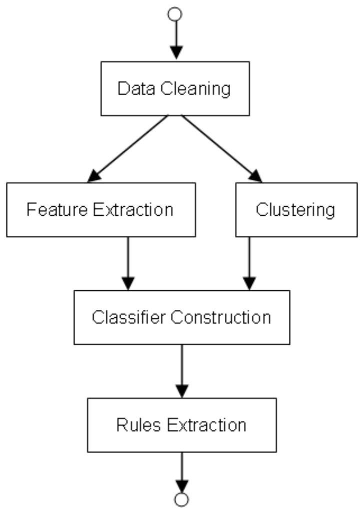

Fig. 6 Two Clusters

A data cleaning procedure filters out inapplicable patient records. For example, patient with only one record is impossible to be analyzed. A feature extraction component transforms the data from the format of matrix-in-matrix to the desirable one with minimum information loss for the purpose of constructing classifiers. A K-means clustering mechanism is applied to produce ideal number of bins for a decision attribute. Decision tree classifiers are then to be constructed for rules extraction.

3 Experiments and Results

Tools used in this research include Microsoft SQL Server text mining package and SAS clustering functions. Types of learning desired for the sparse data in the database D 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. Decision tree study was performed using C4.5 [8], 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 [10]. 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 patient recovery by loudness discomfort levels, new temporal features and problem types.

We performed experiments for three different tasks: Experiment type I explored Tinnitus treatment records of 253 patients and applied 126 attributes to investigate the association between treatment factors and recovery; Experiment type II explored 229 records and applied 16 attributes to investigate the nature of tinnitus with respect to hearing measurements; Experiment type III compared the recovery rate among different category of patients. All classifiers were 10-fold cross validation with a split of 90% training and 10% testing.

3.1 Experiment type I

Total score was used to represent the difference of overall symptom level between the initial visit and the last visit as a decision attribute, where the data domain contains over two hundred different integer values and therefore not suitable for traditional rule-extraction algorithms. Thus, the authors applied K-means clustering techniques to discretize the domain. Intuitively, total score cannot be applied to patients with only one record, and all such patient records thus were removed from the database D. Due to the limitation of the dataset size, the authors tested up to five bins to maintain a desirable number of supports. In Table 1, we observed a significant gap of the average cluster distance between three-bin clustering and four-bin clustering, which indicated that, regarding to the attribute of total score itself, the four-bin clustering was the most efficient one among these experiments. We also observed that in two-bin clustering, the cutting point is at the score of 20, which coincidences the clinical doctor's empirical cut value for a significant recovery.

Table 1 Average Distances

2 Clusters	3 Clusters	4 Clusters	5 Clusters
11.6	10.2	7.0	6.7

In Table 2, the cutting points are listed by the ascending order of total score. Tree classifiers had been constructed based on the above discretization methods and produced interesting rules with different semantics of the decision attribute. This paper focuses on discussing the rules with precision and recall close to and above 60% and a support threshold of seven.

Fig. 7 shows the percentage of new temporal feature rules, where the dark regions represent rules having new temporal features and the light pattern regions stand for rules having no new temporal features. We observed that four-bin clustering

Table 2 Cutting Points

2 Clusters	3 Clusters	4 Clusters	5 Clusters
$(-\infty, 20)$	$(-\infty, -18)$	$(-\infty, -18)$	$(-\infty, -42)$
$[20, +\infty)$	$[-18, 28)$	$[-18, 10)$	$[-18, -14)$
$-$	$[28, +\infty)$	$[10, 36)$	$[-14, 10)$
$-$	$-$	$[36, +\infty)$	$[10, 40)$
$-$	$-$	$-$	$[40, +\infty)$

(with optimal data-driven cutting points) had the highest percentage of new temporal features.

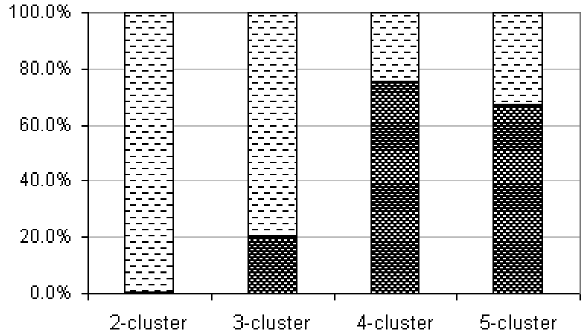

Fig. 7 Rules Having New Temporal Feature vs. Rules Having no Temporal Features

The rest of this section will explain the resultant rules respectively.

3.2 Two-bin Clustering

Tree classifiers were constructed based on the two-bin clustering method, which resulted in a Boolean decision attribute. The semantic meaning of the decision attribute happened to be the same as empirical one from the clinic doctors.

Fig. 8 shows the two-bin clustering, where the total score can be interpreted as either ”significant” or ”insignificant”. In this experiment, we collected the following rules:

Rule 1. *When the total visit is not more than two times, the average WN of the right ear is smaller than 30, and the average right ear sensitivity degree is greater than 90, the recovery tends to be insignificant*.

Rule 2. *When the instrument model is Tr-COE, the recovery tends to be insignificant*.

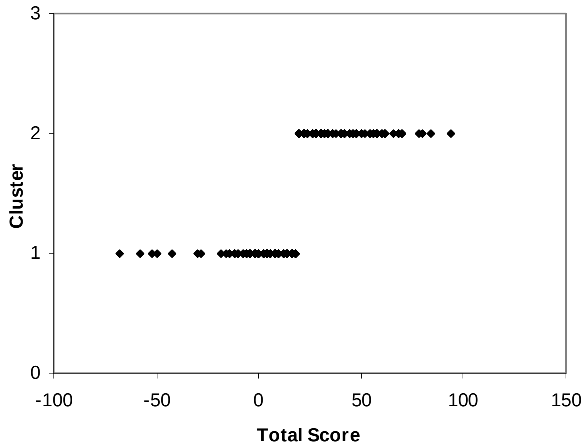

Fig. 8 Two Clusters

Rule 3. *When the problem in the last visit is Tinnitus and the standard deviation of R4 of the right ear is smaller than 5.77, the recovery tends to be insignificant.*

Rule 4. *When the standard deviation of the right ear Tinnitus loudness is smaller than 5.66, the recovery tends to be insignificant.*

Rule 5. *When the instrument model is ITE, the recovery tends to be significant.*

Table 3 For Rules in the Test by Two-Bin Clustering

Rule	Precision (%)	Recall (%)	Support
1	95.8	89.4	23
2	80.0	64.5	8
3	70.6	63.4	12
4	72.7	61.1	8
5	86.7	69.0	13

3.2.1 Three-bin Clustering

Three clusters were generated in this experiment, where the semantic meaning of ”significant” is slightly more restricted than the empirical definition by the clinic doctors. The cutting points of clustering pattern were approximately symmetric, which were complied the common sense of ”neutral”.

Fig. 9 shows the three-bin clustering, where the total score can be interpreted as either ”significantly improved”, ”neutral”, or ”significantly worse”. In this experiment, we collected the following rules:

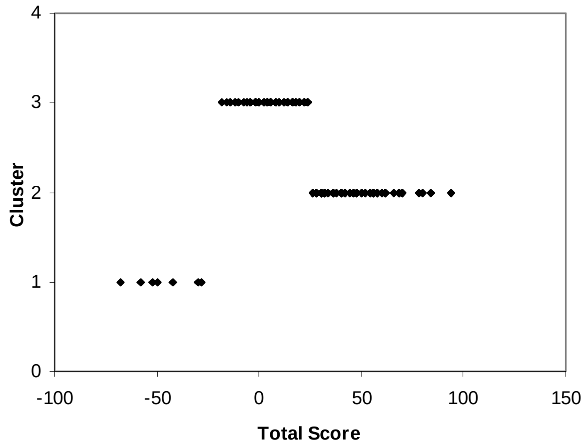

Fig. 9 Three Clusters

Rule 6. *When the average loudness discomfort level-50 of the right ear is less than 74.5, the recovery of a patient tends to be neutral.*

Rule 7. *When the weighted spread of the WN of the left ear is less than 0.29, the recovery of a patient tends to be neutral.*

Rule 8. *When the instrument model is Tr-COE, the recovery of a patient tends to be neutral* (complying Rule 2).

Rule 9. *When the standard deviation of the right ear Tinnitus loudness is smaller than 5.66, the recovery of a patient tends to be neutral* (complying Rule 4).

Rule 10. *When the instrument model is ITE, the recovery tends to be significantly improved* (same as Rule 5).

Table 4 For Rules in the Test by Three-Bin Clustering

Rule	Precision (%)	Recall (%)	Support
6	100	93.0	19
7	73.0	59.3	92
8	80.0	64.5	8
9	75.0	61.1	9
10	87.5	69.0	14

3.2.2 Four-bin Clustering

The four-bin clustering had a larger range for non-insignificant recovery and split the range into two categories. The semantics of category three differed from the doctor's empirical definition of significant, which included both significant and neutral. Assuming that there is an approximately even distribution in each cluster, based on the cutting point, the possibility of significant over neutral is about 1:1.

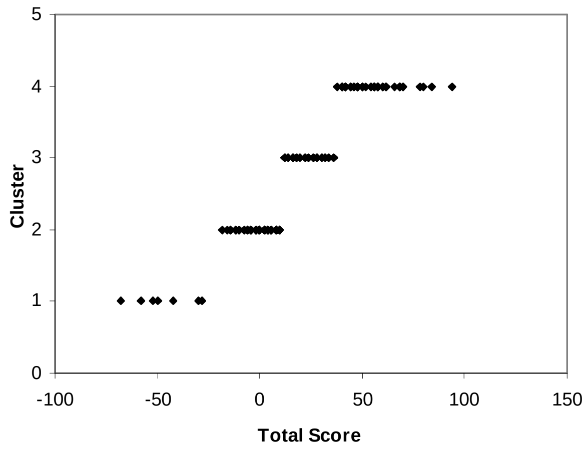

Fig. 10 Four Clusters

Fig. 10 shows the four-bin clustering, where the total score can be interpreted as either ”extensively improved”, ”significantly improved or neutral”, ”neutral”, or ”significantly worse”. In this experiment, we collected the following rules:

Rule 11. *When the average loudness discomfort level-50 of the right ear is less than 74.5 and total number of visits is not 2, the recovery of a patient tends to be neutral* (similar to Rule 6).

Rule 12. *When the symptoms in last visit include both Tinnitus and hearing loss, the weighted centroid of the Tinnitus pitch matched from the left ear is greater than 0.26, and the weighted spread of the threshold of hearing for matching sound for the left ear is not greater than 0.25, the recovery of a patient tends to be significantly improved or neutral*.

Rule 13. *When the symptoms in last visit include Tinnitus, hyperacusis, and hearing loss, the weighted centroid of the Tinnitus pitch matched from the left ear is not greater than 0.125, and the weighted centroid of the loudness discomfort level-8000 of the right ear is greater than 0.64, the recovery of a patient tends to be extensively improved*.

Rule 14. *When the instrument type is ITE, and the weighted spread of the discrimination level of the left ear is not greater than 0.36, the recovery of a patient tends to be extensively improved* (complying Rule 5 and Rule 10).

Table 5 For Rules in the Test by Four-Bin Clustering

Rule	Precision (%)	Recall (%)	Support
11	81.2	70.2	13
12	100.0	88.2	11
13	100.0	82.0	7
14	73.3	53.4	11

3.2.3 Five-bin Clustering

Like the four-bin clustering, the five-bin clustering had a larger range for non-insignificant recovery and split the range into two categories. The semantics of category four differed from the doctor's empirical definition of significance, which included both significant and neutral. Assuming that there is an approximately even distribution in each cluster, based on the cutting point, the possibility of significant over neutral is about 2:1.

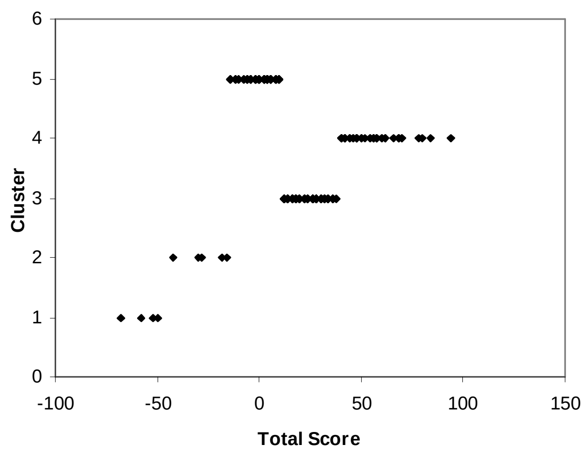

Fig. 11 Five Clusters

Fig. 11 shows the five-bin clustering, where the total score can be interpreted as either "extensively improved", "significantly improved or neutral", "neutral", "significantly worse", or "extensively worse". In this experiment, we collected the following rules:

Rule 15. *When the symptoms in the last visit include both Tinnitus and hearing loss, the total number of visits is not 2, and the average loudness discomfort level-8000 of the right ear is greater than 101, the recovery of a patient tends to be significantly improved or neutral.*

Rule 16. *When the total number of visits is not 2 and average R2 of the right ear is not more than 25, the recovery of a patient tends to be significantly improved or neutral.*

Rule 17. *When the symptoms in last visit include both Tinnitus and hearing loss, the weighted centroid of the Tinnitus pitch matched from the left ear is greater than 0.26, and the weighted spread of the threshold of hearing for matching sound for the left ear is not greater than 0.19, the recovery of a patient tends to be significantly improved or neutral.*

Rule 18. *When the total number of visits is greater than 4, the symptoms in last visit include both Tinnitus and hearing loss, and the weighted centroid of the L12 of the left ear is greater than 0.59, the recovery of a patient tends to be significantly improved.*

Rule 19. *When the symptoms in last visit include Tinnitus, hyperacusis, and hearing loss, the weighted centroid of the tinnitus loudness of the left ear over the sound level threshold of the right ear is not positive, and the weighted spread of the discrimination of the left ear is not greater than 0.15, the recovery of a patient tends to be extensively improved.*

Rule 20. *When the symptoms in last visit include both Tinnitus, and hearing loss, the weighted centroid of L12 of the left ear is not greater than 0.59, and the weighted centroid of the loudness discomfort level for Tinnitus pitch of the left ear is not greater than 0.11, the recovery of a patient tends to be neutral.*

Table 6 For Rules in the Test by Five-Bin Clustering

Rule	Precision (%)	Recall (%)	Support
15	86.7	75.7	13
16	83.3	70.0	10
17	100.0	88.2	11
18	91.7	79.4	11
19	88.9	73.1	8
20	78.6	66.2	11

3.3 Experiment type II

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 related to decreased sound tolerance as indicated by hyperacusis or dislike of sound-misophonia, and phonophobia - 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. Decreased values of loudness discomfort level parameters are not necessary for decreased sound tolerance but they are always lower in case of hyperacusis. In this research the relationship between loudness discomfort level parameters and decreased sound tolerance was investigated. The temporal feature of problems was used as the decision attribute to represent the most often symptoms in all visits of a patient. The value of this feature may be a symptom or a combination of symptoms, where the order in the combination implies the importance of individual symptoms. For example, "THL" means that Tinnitus (T) is the most important symptom and that Hyperacusis (H) is the second most important one and that Hearing Loss (L) is the least important one for a patient. Based on the same thresholds as those in the Experiment I, we collected the following rule:

Rule 21. *When the loudness discomfort level-50 of the right ear is between 19 and 40, Tinnitus tends to be a minor symptom for a patient most time.*
The support of the rules is 27, the precision is 100.0%.

3.4 Experiment type III

Scatter plot analysis in Fig. 12 shows that when recovery rate R is compared to patient category, patient category 4 has a smaller rate of recovery value possibly indicating slower or reduced treatment success than that of other patient categories.

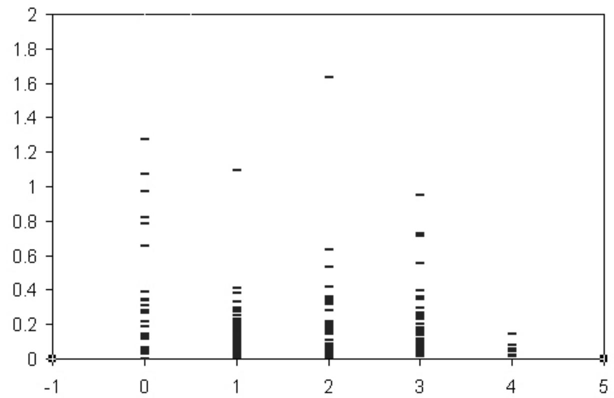

Fig. 12 Scatter plot for the LDL parameters.

4 Conclusion

TRT is a complex treatment process, which generates high volume of matrix data over time: some attributes have relatively stable values while others may be subjective to change as the doctors are tuning the treatment parameters as well as the symptoms of patients are altering. Understanding the relationships between and patterns among treatment factors helps to optimize the treatment process. The authors investigated new features to describe the subtle changes of data, whose domain spans from different disciplines, such as otology as well as audiology for the purpose of classification tree construction. Different clustering methods have been applied to quantize scores into decision attributes. The cutting point of two-bin clustering that was found in the experiment supported the empirical opinion of the clinical doctors about significant recovery. We observed that greater number of bins tended to increase both the precision and recall of the same resultant rules and that rules based on different clustering methods had no conflictions to each other.

Interesting rules about the relationship recovery against symptoms, audiological therapy parameters, and other factors were revealed during the experiments. Also, during the experiments we observed that the new temporal features for continuous data gave more interesting rules especially for extensively improved recovery cases, which confirmed our assumption that carefully designed temporal features may contribute to a better representation of the characteristics of subtle changes over time. Due to the fact that over 77.36% of the patient body in the database has positive recovery results, we may not have enough examples to learn rules about negative cases.

The paper presented an initial research on data mining of medical data about TRT. Expending the patient records will increase the precision as well as the total number of rules. More temporal features for both categorical attributes and continuous attributes shall be explored. It is also interesting to explore the efficient, economic, and independent group of treatment factors to reduce the high expense of the TRT treatment.

References

1. Baguley, D.M.: What Progress Have We Made With Tinnitus, Acta Oto-Laryngologica (Nov.) 556, 4-8 (2006)
2. Heller, M.F., Bergman, M.: Tinnitus in normally hearing persons, Ann Otol, Vol. 62, 73-83 (1953)
3. Henry, J.A., Jastreboff, M.M., Jastreboff, P.J., Schechter, M.A., Fausti, S.: Guide to Conducting Tinnitus Retraining Therapy Initial and Follow-Up Interviews, Journal of Rehabilitation Research and Development, Vol. 40, No. 2, March/April, 159-160 (2003)
4. Jastreboff, P.J.: Tinnitus as a phantom perception: theories and clinical implications. In Vernon, J. and Moller, A. R. (Ed.), Mechanisms of Tinnitus, Boston, MA: Allyn and Bacon, 73-94 (1995)

5. Jastreboff, P.J., Hazell, J.W.P.: Tinnitus Retraining Therapy - Implementing the Neurophysiological Model, Cambridge University Press (2004)
6. Povinelli, R.J., Xin Feng, X.: Temporal Pattern Identification of Time Series Data using Pattern Wavelets and Genetic Algorithms, in Proceedings of Artificial Neural Networks in Engineering, 691-696 (1998)
7. Powers, D., Yu, X.: Statistical Methods for Categorical Data Analysis, Academic Press (1999)
8. Quinlan, J.R.: C4.5: Programs for Machine Learning San Mateo, CA. Morgan Kaufmann (1993)
9. Waldon, M.G.: Estimation of Average Stream Velocity, Journal of Hydraulic Engineering 130(11), 1119-1122 (2004)
10. Witten, I.H., Eibe, F.: Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann Publishers (2005)
11. Zhang, X., Ras, Z.W.: Differentiated Harmonic Feature Analysis on Music Information Retrieval for Instrument Recognition, in Proceedings of IEEE International Conference on Granular Computing, May 10-12, Atlanta, GA., 578-581 (2006)
12. Zhang, X., Ras, Z.W., Jastreboff, P.J., Thompson,P.L.: From Tinnitus Data to Action Rules and Tinnitus Treatment, Proceedings of 2010 IEEE Conference on Granular Computing, Silicon Valley, CA, IEEE Computer Society, 620-625 (2010)
13. Thompson,P.L., Zhang, X., Jiang, W., Ras, Z.W., Jastreboff, P.J., : Mining Tinnitus Database for Knowledge, in Data Mining and Medical Knowledge Management: Cases and Applications, (Eds: P. Berka, J. Rauch, D. Zighed), IGI Global, 293-306 (2009)
14. Nahm, U.Y.: Text Mining with Information Extraction, Ph.D. thesis, Department of Computer Sciences, University of Texas at Austin, August (2004)