Krajowa Konferencja Zastosowań Matematyki w Biologii i Medycynie

Łochów, 23rd–27th September 2014

EMERGENCY PHONE CALL FEATURES ASSOCIATION ANALYSIS

Joanna Grzybowska, Paweł Jaciów and Mariusz Ziółko

Faculty of Computer Science, Electronics and Telecommunications
AGH University of Science and Technology, al. Mickiewicza 30, 30-059 Kraków, Poland
gjoanna@agh.edu.pl, jaciow@student.agh.edu.pl, ziolko@agh.edu.pl

ABSTRACT

Association rule learning was used to establish associations between emergency phone call features. 27 features were analyzed and a large set of rules was obtained as a result. To reduce the number of rules, thresholds for four parameters, calculated for each obtained rule, were fixed. According to the values of five parameters per rule, the rules were sorted in order to select the most important ones. Moreover, we applied bonds on feature positions in the rules. As a result, a set of 437 rules was obtained. The best obtained sets are presented in this paper. The results of our subjective evaluation of the most interesting rules are also presented.

INTRODUCTION

Knowledge Discovery in Databases (KDD) is the process of finding patterns in large data sets in order to extract useful information and learn new aspects of a subject of research. It includes the steps of data selection, preprocessing, transformation, data mining, interpretation and evaluation. Data mining is thus a more specific term – it is an application of an algorithm for pattern extraction [1]. Methods used in the KDD process derive from statistics, machine learning and artificial intelligence. This technique is used in fields such as business, marketing, science, engineering, medicine, bioinformatics, fraud detection and homeland security. The group of application domains is vast.

KDD can be used for a range of purposes such as regression, clustering, and classification. It can also be used for association rule learning, which focuses on discovering important relationships between variables at multiple concept levels. Agrawal et al. [2] introduced association rules that are being used to learn the regularities between purchased products in supermarkets (Market Basket Analysis) in order to determine which products customers are likely to buy together and to predict typical customer behavior. Although association rules are most commonly used for such analysis, they are also useful for finding patterns in many different types of data [3], including our callers database.

The aim of this paper is to investigate the associations of the characteristics of emergency phone calls and to interpret and evaluate the patterns obtained. The results of this analysis can help us determine the most likely profile of a phone call and therefore help in the automatic detection of the speaker's characteristics.

In the following sections, the set of recordings, analyzed call features and their distributions are presented. The applied association rules learning algorithm and the resulting best rules are described. In the last section, potential future applications are proposed together with overall conclusions.

COLLECTING AND PREPARING THE DATA

Our analysis utilizes emergency phone caller data. These are real-life recordings collected by the AGH Digital Signal Processing Group with permission from the Małopolska Emergency Call Center. The corpus is introduced by Witkowski *et. al.* in [4]. The database contains 3 306 recordings, each described with a string of characters that corresponds to a specific set of speaker features perceptually noticed by researchers while annotating the recordings.

Data on gender, age group, speech rate, emotional state with its intensity, intoxication, conversation style, characteristic expressions, voice pathology and environment (acoustic background) was obtained. Additionally, a distinction was made between working days and weekends, days and nights, and shorter and longer than average conversation time.

The number of features describing each phone call was reduced to $N = 27$. We achieved this by: grouping particular emotions into general categories such as neutral, negative and positive; unifying different level of intoxication as just 'intoxicated'; grouping types of voice pathology (pathology in respiratory system or vocal tract) into one class; and discarding features with a negligibly low number of occurrences (less than approx. 1% of all recordings). Weekends were defined as from 10 p.m. on Friday to the end of Sunday, night was defined as from 10 p.m. to 6 a.m. the next day, and long call duration is longer than 49 seconds (the average).

For further processing purposes, we stored the data in a sparse matrix. The matrix, showing some of the feature associations, is presented in Fig. 1. We counted the occurrences of each feature for both genders. Fig. 2 presents this statistical distribution. It can be seen, for example, that male callers are intoxicated and speak chaotically significantly more frequently than female callers.

SEARCHING FOR ASSOCIATION RULES

We performed association analysis to find interesting relationships hidden in our dataset. The aim was to check how frequently some speaker features occur alongside other features in emergency call recordings. The rule follows the format $Head(H) \Rightarrow Body(B)$. *Head* is the condition that needs to be met in order to trigger the rule, and *Body* is the expected result of meeting that condition. For example, a rule with $H = \{\text{man}, \text{weekend}, \text{night}\}$ and $B = \{\text{intoxicated}\}$ can be interpreted as "if a man calls at night at the weekend, he will likely be intoxicated". The concepts of confidence and support identify how strong the found rule is. Let $F = \{f_1, f_2, \ldots, f_n\}$ be a set of n features identifying each phone call. Let $D = \{r_1, r_2, \ldots, r_m\}$ be a set of m recordings called the database. Each recording contains a subset of the features in F. Let the sets of features be called feature-sets. A rule is an implication $H \Rightarrow B$, where $H, B \subseteq F$ and $H \cap B = \emptyset$ [2].

Support

$$\sup(H \cup B) = \frac{\#m(H, B)}{m}$$

is the fraction of recordings m that contain all features from both the head and body of the rule (measures how frequently a rule occurs in the data).

$$\mathrm{conf}(H \Rightarrow B) = \frac{\#m(H, B)}{m(H)} = \frac{\sup(H \cup B)}{\sup(H)}$$

is the ratio of recordings m where features from both the head and body of the rule appear in recordings to the number of recordings that contain features from the head (measures the predictive power of a rule). This parameter tells us how often the presence of features from the head of the rule results in the presence of features from its body in recordings.

We also considered three additional parameters: lift, conviction, and Laplace measure. Lift

$$\mathrm{lift}(H \Rightarrow B) = \frac{\mathrm{conf}(H \Rightarrow B)}{\sup(B)} = \frac{\sup(H \cup B)}{\sup(H) \times \sup(B)}$$

is a support of features from the head and body of a rule to the support expected if H and B were independent. Lift ranges from 0 to $+\infty$ and if it is close to 1, H and B are considered independent and thus the rule is not interesting. The greater this value is, the more information is provided on the body by the head — the rule body and rule head appear together more frequently than expected [5]

Conviction

$$\mathrm{conv}(H \Rightarrow B) = \frac{1 - \sup(B)}{1 - \mathrm{conf}(H \Rightarrow B)}$$

of a rule is understood as the frequency of the rule making an untrue prediction (that H occurs without B). Conviction measures the degree of implication of a rule. Unlike lift, conviction is sensitive to rule direction. Conviction ranges from 0.5 to $+\infty$, it is infinite for logical implications ($\mathrm{conf} = 1$), and the value close to 1 indicates that H and B are independent. In conviction, unlike confidence, the support of both the head and body are considered.

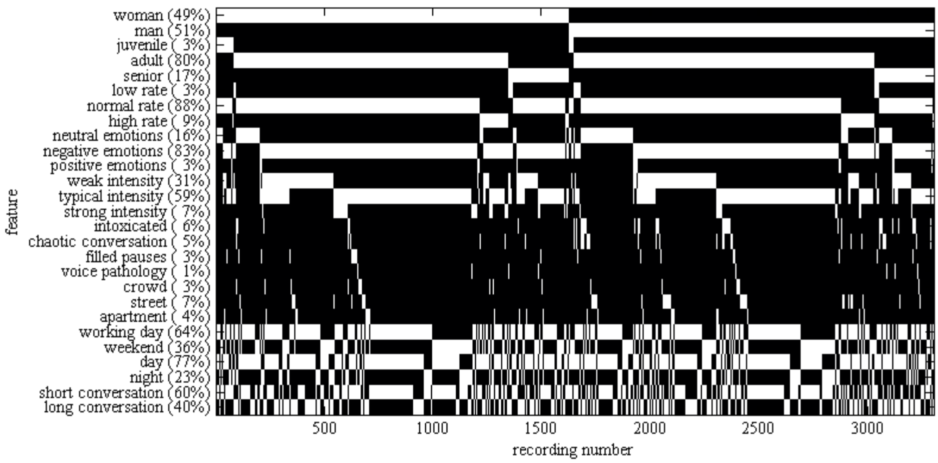

Figure 1. A sparse matrix of the data sorted by features from top to bottom. White color indicates the presence of a feature.

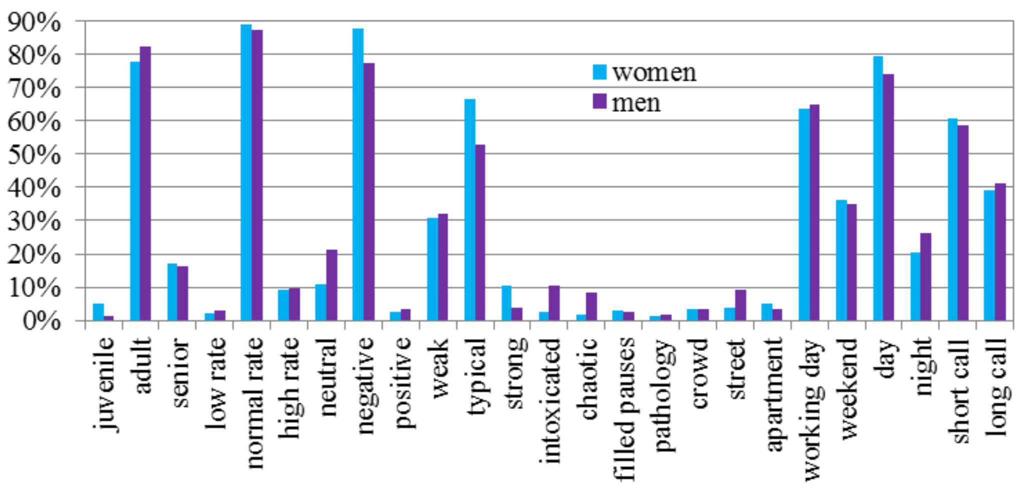

Figure 2. Statistical distribution of features for women and men.

Laplace measure

$$\mathrm{lapl}(H \Rightarrow B) = \frac{\sup(H \cup B) + 1}{\sup(H) + 2}$$

is a confidence of a rule, although it focuses more on the support of H. As the support of H decreases, Laplace measure also decreases. Values of Laplace measure range from 0 to 1 [5].

Looking for good association rules comes down to performing two steps: finding sets of features with support not less than the given support threshold (finding frequent feature-sets), and selecting combinations with other parameter values greater or equal to their respective thresholds. We decided to also apply lift and conviction threshold values. Based on the parameters described, the Apriori Algorithm developed by Agrawal *et al.* [6], [7] performs the following steps:

1) F_1 = {one-feature-sets} // Count feature occurrences to determine frequent // one-feature-sets F_1;
2) for $(n = 2; F_{n-1} \neq \mathit{empty}, n = n + 1)$ // Then repeat until the last generated set is empty;
3) $C_n = \{x \cup \{y\} | x \in F_{n-1} \wedge y \in \bigcup F_{n-1} \wedge y \notin x\}$; // Based on F_{n-1} generate candidate // n-feature-sets C_n;
4) for recording $r \in D$ // Next, scan the database...;
5) $C_r = \{c | c \in C_n \wedge c \subseteq r\}$;
6) for candidate $c \in C_r$ // ... and count support of candidates...;
7) count[c] = count[c] +1;
8) $F_n = \{c | c \in C_n \wedge \mathrm{count}[c] \geq \mathit{treshold}\}$ // ... to find frequent n-feature-sets F_n;
9) Result $= \bigcup_n F_n$.

After that, given the frequent feature-sets, association rules are generated and evaluated against three parameter thresholds (rules with parameters lower than their thresholds are discarded).

FINDING ASSOCIATION RULES

As a result of the algorithm described in the previous section, with only one threshold for support set to 0.008, a total of 227 208 rules were obtained. Then, setting thresholds to 0.6 for confidence, 0.008 for support, 1 for lift and 1 for conviction, we reduced the number of rules to 11 863. The distribution of these rules (support-confidence dependency with lift color indication for each rule) is presented in Fig. 3. [!htb]

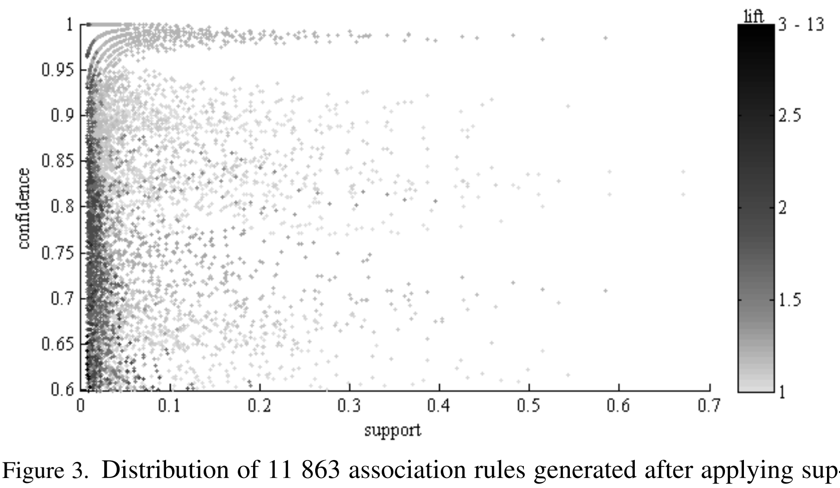

Figure 3. Distribution of 11 863 association rules generated after applying support, confidence, lift and conviction thresholds.

The rules were then sorted to obtain the best ones. Lift was chosen as the most important factor – the features were sorted with the descending lift value. The other parameters were sorted in the

Table 1. One best rule for each body feature after fixing thresholds.

Head	→ Body	sup	conf	lift	conv	lapl
neutral emotions, intoxicated, day	→ chaotic conversation	0.01	0.66	12.58	2.78	0.36
man, senior, normal speech rate, chaotic conversation, day	→ neutral emotions	0.01	0.64	3.99	2.35	0.34
voice pathology	→ senior	0.01	0.61	3.66	2.15	0.50
man, adult, normal speech rate, negative emotions, street, working day, long conversation	→ weak intensity	0.01	0.76	2.42	2.88	0.34
woman, normal speech rate, weak intensity, apartment	→ long conversation	0.01	0.93	2.31	8.65	0.34
senior, chaotic conversation, long conversation	→ man	0.01	1.00	1.97	Inf	0.40
juvenile, normal speech rate, negative emotions, short conversation	→ woman	0.01	0.91	1.86	5.76	0.34
woman, strong intensity, night	→ weekend	0.01	0.66	1.85	1.89	0.37
typical intensity, filled pauses, working day, day	→ short conversation	0.01	1.00	1.68	Inf	0.34
woman, negative emotions, filled pauses	→ typical intensity	0.01	0.93	1.57	5.95	0.35
neutral emotions, chaotic conversation, day, long conversation	→ working day	0.01	0.97	1.51	11.42	0.35
senior, neutral emotions, chaotic conversation, working day	→ day	0.01	1.00	1.30	Inf	0.37
high speech rate, street	→ adult	0.01	1.00	1.25	Inf	0.47
strong intensity	→ negative emotions	0.07	1.00	1.21	Inf	0.52
juvenile, neutral emotions	→ normal speech rate	0.01	1.00	1.14	Inf	0.46

following: conviction, Laplace measure, confidence, and support. The rules for 15 single body features (out of 27) were present in the set of almost 12 000 rules. One best rule for each of the features appearing in the body of a rule is presented in Table 1. In the next step of the association analysis, we have chosen the features to be fixed in the head and in the body of a rule. The characteristics which may appear in the head of a rule are gender, speech rate, manner of speaking, pathological speech, weekend/working day, day/night, and conversation length. In the body of a rule the following characteristics may appear: age, emotional state, intensity of emotions, intoxication, and acoustic background. This way the number of rules was reduced to 437, obtained for four different single body features. The best rules for these features are presented in Table 2.

As well as finding rules as described above, we also looked for interesting ones:

- If emotion intensity is strong, it is a negative emotion (100%);
- Callers whose intensity of emotions is strong are more frequently women (71%);
- Chaotically-speaking callers are also intoxicated (44%);
- Emotionally-neutral callers speak chaotically more frequently than others (5 times more);
- Callers who speak chaotically almost never express positive emotions (98.3%);
- Intoxicated people tend to call more often at night than during a day (twice as frequently);
- Intoxicated callers tend to have long calls (62%);
- Juvenile callers are mostly women (77%);
- Calls at night are more common at weekends than during working days (twice as common).

Table 2. One best rule for each body feature after applying bonds on features positions in rules.

Head	**→ Body**	**sup**	**conf**	**lift**	**conv**	**lapl**
voice pathology	→ senior	0.01	0.61	3.66	2.15	0.50
woman, filled pauses, short conversation	→ typical intensity	0.01	0.90	1.52	4.26	0.36
man, high speech rate, working day, short conversation	→ adult	0.01	0.98	1.22	8.77	0.34
filled pauses, working day, short conversation	→ negative emotions	0.01	1.00	1.21	Inf	0.35

CONCLUSIONS AND FUTURE PLANS

Using association rules enabled us to reveal correlations in the dataset without any prior knowledge of what patterns to look for. The set of obtained rules was reduced and as a result it was possible to subjectively assess and evaluate the findings. The set of rules generated can also be filtered in a different manner, depending on the particular researcher's area of interest, e.g. rules applying to a specific type of emotion can be extracted.

The best rules were chosen according to specified criteria. However, depending on the aim of the analysis, other selection strategies could also be applied, for example with an emphasis on strong rules, which are those with both high support and confidence [3].

The set of features analyzed in this study is very similar to the set of features in a system for caller identification by voice [4]. The relationships discovered in this study can therefore be applied to the abovementioned caller identification system to influence the performance of automatic recognition. As a new component of the system's knowledge base, the discovered rules may be used to predict features that could not be explicitly detected automatically. The rules may also help us set up confidence thresholds for the detection of predicted features in the system. Future work aims at implementing the knowledge about the relationships of the features to obtain complete caller profiles. While reducing the number of rules by setting the bonds on feature position, the following features were chosen for the body of a rule: age, emotional state, intensity of emotions, intoxication, and acoustic background. This choice was driven by the possibility of applying these rules in an automatic system for caller identification by voice in the future.

Future work may also include an analysis of the 242 recordings of 30 different callers who contacted the Emergency Call Center multiple times. Such research would enable us to perform analyses that would be more finely adjusted to the particular speaker and would result in more personalized rules.

ACKNOWLEDGEMENTS

The project was supported by the National Research and Development Center granted by decision 072/*R*/*ID*1/2013/03.

REFERENCES

[1] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth: *From Data Mining to Knowledge Discovery in Databases*, AI MAGAZINE, Association for the Advancement of Artificial Intelligence (1996), Retrieved 17 December 2008.

[2] R. Agrawal, T. Imielinski, and A. Swami: *Rules between Sets of Items in Large Databases*, Proceedings of the 1993 ACM SIGMOD Conference, Washington DC, 1993, pp. 207–216.

[3] B. Lantz: *Machine Learning with R*, Packt Publishing, Birmingham, 2013, Chapter 8.

[4] M. Witkowski, M. Igras, J. Grzybowska, P. Jaciów, J. Gałka, and M. Ziółko: *Caller Identification by Voice*, Proceedings of the XXII Annual Pacific Voice Conference, Kraków 2014.

[5] P.J. Azevedo and A.M. Jorge: *Comparing Rule Measures for Predictive Association Rules*, ECML '07 Proceedings of the 18th European conference on Machine Learning, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 510–517.

[6] R. Agrawal and R. Srikant: *Fast Algorithms for Mining Association Rules*, Proceedings of the 20*th* VLDB Conference, Santiago, 1994, pp. 487–499.

[7] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A.I. Verkamo: *Fast discovery of association rules*, Advances in knowledge discovery and data mining, Menlo Park 1996, pp. 307–328.