ANOVA Based Approach for Efficient Customer Recognition: Dealing with Common Names
Morteza Saberi, Zahra Saberi

To cite this version:
Morteza Saberi, Zahra Saberi. ANOVA Based Approach for Efficient Customer Recognition: Dealing with Common Names. 4th IFIP International Conference on Artificial Intelligence in Theory and Practice (AI 2015), Oct 2015, Daejeon, South Korea. pp.64-74, 10.1007/978-3-319-25261-2_6. hal-01383957

HAL Id: hal-01383957
https://inria.hal.science/hal-01383957
Submitted on 19 Oct 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License
ANOVA Based Approach for Efficient Customer Recognition: Dealing with Common Names

Morteza Saberi  
School of Business, UNSW Canberra BC 2610 Australia

Zahra Saberi  
School of Industrial Engineering, University of Tehran, Iran

Abstract  
This study proposes an Analysis of Variance (ANOVA) technique that focuses on the efficient recognition of customers with common names. The continuous improvement of Information and communications technologies (ICT) has led customers to have new expectations and concerns from their related organization. These new expectations bring various difficulties for organizations’ help desk to meet their customers’ needs. In this paper, we propose a technique that provides the most beneficial information to the Customer service representative that will assist in the efficient recognition of the customer. The proposed algorithm determines which features of a customer should be asked that would result in his/her prompt recognition. Moreover, to have a clean database, the framework uses the features of customers for which a standard format is available such as street address, month of birth etc. We evaluate our algorithm on synthetic dataset and demonstrate how we can recognize the right customer in the optimum manner.

Keywords—Contact centres, customer recognition, Customer common name,

I. INTRODUCTION  
Customer relationship management (CRM) is a framework for managing a company’s interactions with current and future customers [1, 2]. In this highly competitive world, it is essential for every organization to have an efficient and smart CRM system. One important part of CRM system is contact center which is in direct contact with the customers. Contact center has been termed as the new version of call centers which allow customers to express their queries via different communication channels: telephone, touch-point telephone, fax, letter, email and online live chat [3]. However, for the contact centers to be effective, they need to be able to identify the customers in question from their database efficiently and effectively. To achieve that, in this study a framework is proposed that utilizes the ANOVA technique to assist contact centers with efficient customer recognition which using telephone and online live chat as the main communication channels.

Currently, individuals are in contact with organization via diverse communication channels [4]. This diversity has two impacts: increasing easiness and flexibility of communication and producing dirty data [5]. The first one is beneficial for the customers while the second one can be a risk for the efficient working of a CRM system. Performance of CRM systems is decreased by existence of dirty data in them. Another issue with CRM system is that customers have different IDs from different organizations. It should be noted that they do not normally memorize all of their IDs. From a psychological perspective, customers prefer to be recognized by their name as their identity, not by a customer number or ID number. The feeling of ownership achieved from a name is much greater than that achieved from address, phone number etc. In some cultures, people who have nicknames prefer to use them as their official given names. Recognizing customers using their name is not a difficult task if the same or similar names are not available in the organization database. However, the recognition is difficult when there are common names in the CRM database. As an example, 7.4% of family name in china is ‘Li’ and by considering 1 milliard as the china population it is 74 million individuals with “Li” as the family name [6]

In the current study, we have focused on customer recognition in the case of existence of the common name in CRM database by making use of ANOVA Based Approach. The main aim of the proposed algorithm is to recognize the customer by asking the minimum number of questions with them. The output of the algorithm is a sequence of most informative questions (customer features) that needs to be asked from the customer. It achieves this by using two sources of information: customer
record(s) in organization database and customer interactions (answer) with the CSR (or: information provided by customers in their interaction) The proposed framework links the customer interactions (responses) with the organization database in order to fulfill customer recognition task. The first data source of information (CRM database) can contain noisy and dirty data.

![Diagram](image)

**Figure 1. Problem definition**

The paper is organized as follows: Section II defines the problem along with a discussion on common name issue for contact centers. The framework of ANOVA based approach for efficient customer recognition is explained in Section III.

### II. COMMON NAMES

Individuals use their personal names for communication purposes [6]. A personal name usually includes a given name and a surname (or family name). In some countries and cultures, a middle name is used as part of the personal name. For example, in Brazil the mother’s family name is given to the child as his/her middle name at birth. In some countries, it is common for individuals to have only one name (mononymous) by which they are recognized. It is interesting to note the various origins of given names:

It is evident that various factors such as culture, language, religion, history, translation to a new language, and environment may determine the personal name given to an individual [7]. These fields have been also studied by researchers and are known as Anthroponymy study. Briefly the mentioned various factors lead to produce different variations of personal names.

**Figures.** Some figures are presented here about the statistic of popular and common name. Table 1 shows the top five popular boys and girls name in 2013 in UNSW, Australia. It is clear even within a given state the number of common name is high. This issue is more severe in countries such china. As an example, 7.4% of family name in china is ‘Li’ and by considering 1 milliard as the china population it is 74 million individuals with “Li” as the family name. Some organization such as health care, insurance, telecommunications company that have high number of customer are more faced with the issue of common name.

<table>
<thead>
<tr>
<th>Boys</th>
<th>Number</th>
<th>Girls</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>William</td>
<td>696</td>
<td>Charlotte</td>
<td>613</td>
</tr>
<tr>
<td>Oliver</td>
<td>630</td>
<td>Olivia</td>
<td>551</td>
</tr>
<tr>
<td>Jack</td>
<td>574</td>
<td>Amelia</td>
<td>540</td>
</tr>
<tr>
<td>Noah</td>
<td>555</td>
<td>Chloe</td>
<td>517</td>
</tr>
<tr>
<td>James</td>
<td>510</td>
<td>Mia</td>
<td>517</td>
</tr>
</tbody>
</table>

### III. ANOVA BASED APPROACH FOR EFFICIENT CUSTOMER RECOGNITION

This section presents the general framework of interactive customer feature selection. This framework assists CRM systems with customer recognition when the system faces the common name issue. The proposed framework relies on two integrated modules as shown in Figure 1.

Analysis of Variance (ANOVA) is developed by Fisher to find out whether is it any difference among groups average [7]. We performed ANOVA on customer standard features to find if means of these features are statistically different or not. Null hypothesis can be accepted or rejected by performing ANOVA F-test in a consequence. We use average of feature values as the index for selecting the first question if ANOVA shows no difference among standard features (Null Hypothesis acceptance). Also if the null hypothesis in ANOVA F-test is rejected, multiple pair comparison is performed to find the most informative question. Before presenting the proposed Duncan's multiple range test (DMRT) based algorithm in a formal way it has been explained with the following three examples.

**Example 1.** This example shows how we select the first question when the null hypothesis is rejected. A multiple pair comparison is performed in this case, as shown in Table 2. The number of rejections associated with each feature is highlighted in Table 3. As the maximum number belongs to Street, it has been selected in this example. This maximum number is the reason that Street is the source of rejection in the null hypothesis.

<table>
<thead>
<tr>
<th>Null Hypothesis &amp; Pairwise Comparisons</th>
<th>Decision</th>
<th>Selected Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_{street} = \mu_{street} = \mu_{street}$</td>
<td>reject</td>
<td>Street</td>
</tr>
</tbody>
</table>
Table 3. Associated Number of rejection for customer features in Multiple pair comparison test (Example )

Example 2. Examination of Table 4, as the second example, shows the first question could be determined between Street & Suburb in this case. From statistical viewpoint customers month values are not different (statistically) with Street and suburb and the most informative question should be determined from Street & Suburb. Also number of rejections for these two features is equal as mentioned in Table 5.

Table 4. Multiple pair comparison example with 2 hypothesis acceptances

Table 5. Associated Number of rejection for customer features in Multiple pair comparison test

Example 3. In Table 6, we have the case in which three attribute are nominated after performing multiple pair comparison. As in Table 7 stated the number of rejection for all customer features is equal

Table 6. Multiple pair comparison example with 3 hypothesis rejections

Table 7. Associated Number of rejection for customer features in Multiple pair comparison test

Now we should answer this question: **How we should determine the optimum attribute in case of two or three nominated attributes?**

As formally P-value shows how rejection or acceptance is strong, we use this value to come up to the best question (attribute). We utilize this nature of P-value that **the more P-value distance with confidence levels the more robust statistical rejection.** The detailed mathematical presentation of this approach has been explained later.

**A. Analysis of Variance based algorithm formal definition**

In the previous section we have provided intuitive example that shows how ANOVA approach help us to find the most informative question. We have two main hypothesis regarding ANOVA usage in customer recognition. The null hypothesis assume the statistically equality of all customer profile features average and on the other hand alternative hypothesis assume average of customer profile features are different. If null hypothesis is rejected then we are in need of finding the cause of this rejection. The feature that leads to this rejection (inequality of features average) is the most informative question and should be asked from the customer. Lines 3 to 18 in Figure 1 shows how find the cause of this rejection. When we get rejection the multiple pair comparison test is then performed and based on number of rejection that is associated with each feature the first feature is selected. The detail of selection is highlighted in the algorithm. Also the process of finding sequence of questions has been depicted in next section. In ANOVA based algorithm some functions have been used as their task are listed in Table 8.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Number of rejection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Street</td>
<td>2</td>
</tr>
<tr>
<td>Suburb</td>
<td>2</td>
</tr>
<tr>
<td>Month of Birth</td>
<td>2</td>
</tr>
</tbody>
</table>

### ANOVA based algorithm (A)

**Input:** A set of records with standard features \( A \),
Number of rows \( (A) \) \( m \), ANOVA F-test ANOVA;

**Output:** preferable feature

\( T \) = Distance\( (A) \);

1. \( H \leftarrow \text{ANOVA}(T) \);
2. \( \text{if } (H=1) \)
In the following figures (2–4), procedures of finding out the three functions are explained:

1. **Distance, Max_Distance and Max_Dif_Pvalue.**

   **Figure 1. ANOVA algorithm**

   **Table 8. ANOVA Algorithm functions definition**

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T \leftarrow \text{Distance}(A)$</td>
<td>Return the matrix $T$ that is average distance of correspondent member of $A$ from other members in its column.</td>
</tr>
<tr>
<td>$M \leftarrow \text{Mul}_\text{Comp}(T, a)$</td>
<td>Perform multiple pair comparison on $T$ and returns a matrix $M$ of pairwise comparison results with a confidence level.</td>
</tr>
<tr>
<td>$L \leftarrow N_\text{Rejection}(M)$</td>
<td>Return dataset array $L$ that shows the number of rejection which each feature got in total pairs.</td>
</tr>
<tr>
<td>$z \leftarrow \text{Maximal_Index}(L)$</td>
<td>Returns indices of the maximum values of $L$ in output vector $z$.</td>
</tr>
<tr>
<td>$\text{Max_Distance}(1:z)$:</td>
<td>Returns indices of features which the distance between its string values is the maximum (maximal) in comparison with other features.</td>
</tr>
<tr>
<td>$\text{Max_dif_Pvalue}(M/z)$:</td>
<td>first calculate the difference between p-value and critical value and find index(es) with the maximum(maximal) value(s). The detail about this function has been highlighted in Figure.</td>
</tr>
</tbody>
</table>

   In the following figures (2–4), procedures of finding out the three functions are explained: Distance, Max_Distance and Max_dif_Pvalue.

   **Figure 2. Distance function**

   **Figure 3. Max_Distance function**

   **Figure 4. Max_dif_Pvalue function**

   **B. Generating sequence of questions by using ANOVA algorithm: Formal definition**

   The way of generating sequence of questions by using ANOVA is explained here. As it is stated in line 14 (Figure 5), when we have two remained feature then we select the question by using Max_distance function. Actually when we have two features using ANOVA test is meaningless and we just used simply edit distance metric. The feature with highest distance among its values is selected with this function. To have a better follow up the example has been given here as well that shows how the sequence of questions is generated in ANOVA approach.
Algorithm ANOVA(Customer Name)

Input: X as common name, a set of records with Name(X)
A, List of Common names C, number of customer standard
features n, a set of standard features Z;
Successful customer recognition flag;

Output: Sequence of questions (features) SQ

1 IF Y ∈ C THEN
2 Break
3 Else X ← Y End
4 A ← Records with name X
5 \( \text{flag} \leftarrow 0 \)
6 While \( \text{flag} = 0 \)
7 \( k \leftarrow n - i \)
8 IF \( k>2 \) THEN
9 \( C' \leftarrow \text{ANOVA}(A) \)
10 \( SQ(i) = C'; Z = Z - \{C'\} \)
11 Else
12 \( C' \leftarrow \text{Max_distance}(A|x) \)
13 \( SQ(i) = C' \)
14 End
15 \( \text{flag} \leftarrow \text{Recognition}(A,k',p) \)
16 \( i \leftarrow i + 1 \)
17 End
18 A ← Records with \( SQ(i) \) customer answer
19 IF \( i=n \)
20 Go to 20
21 End
22 Go to 9
23 Ask sequence of other features
24 END
25 Ask sequence of other features
26 Go to 9
27 Return SQ;

Figure 5. ANOVA Based Approach pseudo-code

Figure 6 also depicts flowchart of generating sequence of questions by using ANOVA algorithm.

Figure 6. Generating sequence of questions by using ANOVA algorithm

C. Example

We show how ANOVA based algorithm finds Mary Miller profile which is mentioned in Table 9. As stated in Table 10 by performing ANOVA F-test, Null Hypothesis is rejected and we should do the multiple comparisons test to find the sequence of questions.

Table 9. Generating sequence of questions’ dataset example

<table>
<thead>
<tr>
<th>Customer Name</th>
<th>Street</th>
<th>Suburb</th>
<th>Month of Birth</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Mary Miller</td>
<td>South Guildford</td>
<td>Donnelly River</td>
<td>June</td>
</tr>
<tr>
<td>2 Mary Miller</td>
<td>Hart close</td>
<td>Dural</td>
<td>July</td>
</tr>
<tr>
<td>3 Mary Miller</td>
<td>Neerabup</td>
<td>Doodenanning</td>
<td>July</td>
</tr>
<tr>
<td>4 Mary Miller</td>
<td>Nollamara</td>
<td>Doongin</td>
<td>April</td>
</tr>
<tr>
<td>5 Mary Miller</td>
<td>Noranda</td>
<td>Daadenning Creek</td>
<td>January</td>
</tr>
<tr>
<td>6 Mary Miller</td>
<td>North Beach</td>
<td>Dagger Hills</td>
<td>May</td>
</tr>
<tr>
<td>7 Mary Miller</td>
<td>Stuart</td>
<td>Palmerston</td>
<td>September</td>
</tr>
<tr>
<td>8 Mary Miller</td>
<td>Stuart</td>
<td>Palmerston</td>
<td>October</td>
</tr>
<tr>
<td>9 Mary Miller</td>
<td>Stuart</td>
<td>Palmerston</td>
<td>September</td>
</tr>
<tr>
<td>10 Mary Miller</td>
<td>South Guildford</td>
<td>Donnelly River</td>
<td>November</td>
</tr>
<tr>
<td>11 Mary Miller</td>
<td>Preston</td>
<td>Emu phlains</td>
<td>December</td>
</tr>
<tr>
<td>12 Mary Miller</td>
<td>Preston</td>
<td>Emu phlains</td>
<td>April</td>
</tr>
<tr>
<td>13 Mary Miller</td>
<td>South Guildford</td>
<td>West Ballidu</td>
<td>January</td>
</tr>
</tbody>
</table>

Table 10. Null Hypothesis ANOVA F-test performing multiple comparisons

<table>
<thead>
<tr>
<th>Null Hypothesis</th>
<th>Decision</th>
<th>P-value</th>
</tr>
</thead>
</table>

5
Result of Multiple comparisons test can be obtained from Table 11. Rejection (1) or acceptance (0) and associated p-values are reported in this table. As number of rejections is equal for all three features (Table 12) the P-value is employed to determine the first question. According to the reported values in Table 13, Month is the first question that should asked from John.

As John’s month of birth is July, we come up with the following updated records (Table 14). As we have just two remaining features the next question is selected by using Max_distance function. Examination of Table 15 shows the next question is John’s Street address.

As John’s month of birth is July, we come up with the following updated records (Table 14). As we have just two remained feature the next question is selected by using Max_distance function. Examination of Table 15 shows the next question is John’s Street address.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Street</td>
<td>0.9951</td>
</tr>
<tr>
<td>Suburb</td>
<td>0.9951</td>
</tr>
<tr>
<td>Month of Birth</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Table 14. Updated dataset based on customer answer

<table>
<thead>
<tr>
<th>Street</th>
<th>Suburb</th>
<th>Month of Birth</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Hart close</td>
<td>July</td>
</tr>
<tr>
<td>3</td>
<td>Neerabup</td>
<td>Doodenanning</td>
</tr>
</tbody>
</table>

Table 15. Max_distance function Result

Conclusion

This study present an interactive customer feature selection to deal with common variations in personal names. Three algorithms from different schools of thoughts, information retrieval, machine learning, and statistical analysis, to find the optimum sequence of questions. These algorithms are IDF and Levenshtein based, C4.5 based and ANOVA based. The preferred approach is the one which require the minimum number of customer interactions. About 60000 records as the synthetic data are used to show the applicability of the proposed framework. Febrl software as the data generator used various parameters that these let get more insights from the framework’s operation. The unique features of the proposed framework enable it to: have online responses, deal with common names, improve the cleansing of the database, have faster performance. The preferred algorithm is selected fast and smooth recognition. The framework has been designed so that it improves the cleansing quality of the CRM system database (customer’s data); a clean database improves performance and subsequently leads to better customer satisfaction.

Reference


