ANOVA Based Approch for Efficient Customer Recognition: Dealing with Common Names

Morteza Saberi, Zahra Saberi

ANOVA Based Approch 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.

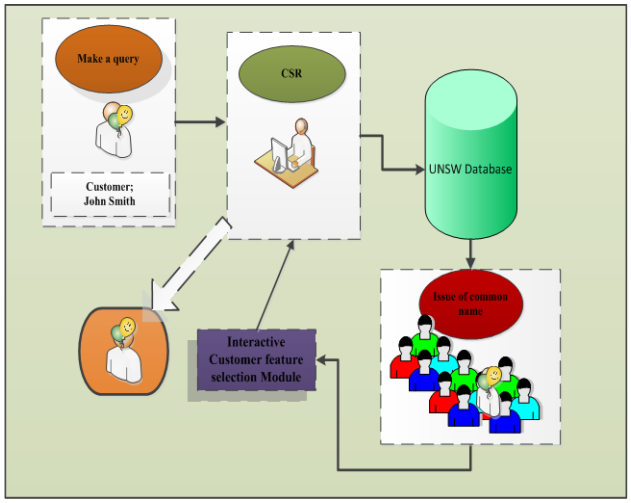

Figure 1. Problem definition

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 1. Popular Baby Names in UNSW 2013

Boys	Number	Girls	Number
William	696	Charlotte	613
Oliver	630	Olivia	551
Jack	574	Amelia	540
Noah	555	Chloe	517
James	510	Mia	517

III. ANOVA BASED APPROCH 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 2. Multiple pair comparison example with 1 hypothesis acceptance

Null Hypothesis & Pairwise Comparisons	Decision	Selected Question
$\mu_{street} = \mu_{suburb} = \mu_{Month}$	reject	

Table 3. Associated Number of rejection for customer features in Multiple pair comparison test (Example)

$\mu_{street} = \mu_{suburb}$	reject	
$\mu_{street} = \mu_{Month}$	reject	
$\mu_{suburb} = \mu_{Month}$	accept	**Street**

Feature	Number of rejection
Street	2
Suburb	1
Month of Birth	1

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

Null Hypothesis & Pairwise Comparisons	Decision	Selected Question
$\mu_{street} = \mu_{suburb} = \mu_{Month}$	reject	
$\mu_{street} = \mu_{suburb}$	accept	
$\mu_{street} = \mu_{Month}$	accept	
$\mu_{suburb} = \mu_{Month}$	accept	**Street, Suburb**

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

Feature	Number of rejection
Street	1
Suburb	1
Month of Birth	0

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

Null Hypothesis & Pairwise Comparisons	Decision	Selected Question(s)
$\mu_{street} = \mu_{suburb} = \mu_{Month}$	reject	**Street, Suburb, Month**
$\mu_{street} = \mu_{suburb}$	reject	
$\mu_{street} = \mu_{Month}$	reject	
$\mu_{suburb} = \mu_{Month}$	reject	

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

Feature	Number of rejection
Street	2
Suburb	2
Month of Birth	2

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.

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*** ← **ANOVA**(**T**);
2 **If (H=1)**

3 $\boldsymbol{M} \leftarrow Mul\_Comp(\boldsymbol{T}, \boldsymbol{\alpha})$ $\boldsymbol{L} \leftarrow N\_Rejection(\boldsymbol{M})$
5 $\boldsymbol{z} \leftarrow Maximal\_Index(\boldsymbol{L})$
6 **If (size(z)<2)**
7 **Return** $Feature_z$; **Else if (size(z)==2)**
8 $\boldsymbol{x} \leftarrow$ **Max_Distance**$(\boldsymbol{T}|\boldsymbol{z})$
9 **Return** $Feature_x$
10 **Else**
12 $x \leftarrow$ max_dif_Pvalue$(M|z)$, $k \leftarrow$ size(x)
14 **If k=1**
15 **Return** $Feature_x$
16 **Else** $y \leftarrow$ **max_distance**$(\boldsymbol{T}|\boldsymbol{x})$, **Return** $Feature_y$
17 End
18 End
19 **Else**
20 $y \leftarrow$ **max_distance**$(\boldsymbol{T}|\boldsymbol{x})$, **Return** $Feature_y$
21 **End**

Figure 1. ANOVA algorithm

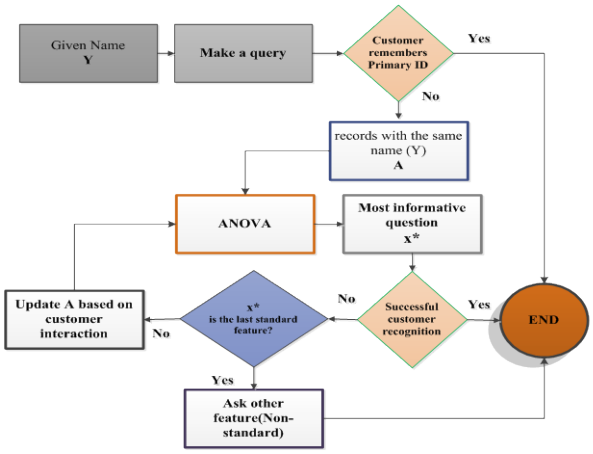

Table 8. ANOVA Algorithm functions definition

function	description
$T \leftarrow$ Distance(A)	*Return the matrix T that is average distance of correspondent member of A from other members in its column*
ANOVA (T)	*Perform ANOVA F-test and determine the acceptance or rejection status of Null Hypothesis. If test reject the null hypothesis then the H is equal to **1**.*
M ←Mul_Comp (T, $\boldsymbol{\alpha}$)	*Perform multiple pair comparison on T and returns a matrix M of pairwise comparison results with α confidence level.*
$\boldsymbol{L} \leftarrow N\_Rejection(\boldsymbol{M})$	Return dataset array L that shows the number of rejection which each feature got in total pairs.
$\boldsymbol{z} \leftarrow Maximal\_Index(\boldsymbol{L})$	Returns indices of the maximum values of L in output vector z.
Max_Distance (T/z):	Returns indices of features which the distance between its string values is the maximum (maximal) in comparison with other features.
Max_dif_Pvalue(M/z):	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.

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

distance$(\boldsymbol{A})$
Input: A set of records with standard features ***A***,
Number of rows (A) ***m,*** Normalized levenshtein distance ***Leven***;
Output: Distance of matrix A member from others
$\boldsymbol{distance}_j = Average_{i \neq h}(Normalized\_Leven(\mathrm{A(i,j)}, \mathrm{A(h,j)})$
Return $\boldsymbol{distance}$**;**

Figure 2. Distance function

Max_distance$(\boldsymbol{A}|\boldsymbol{z})$
Input: A set of records with standard features ***A***,
Number of rows (A) ***m,*** ;
Output: indices of the maximum distances among each feature
For $j \in z$: $(A|z) \leftarrow A - A(:, j)$ **End**
$\boldsymbol{dist}_j =$ **distance**$(\boldsymbol{A}|\boldsymbol{z})$
$j^* \leftarrow \arg\_\max_j(dist_j)$
Return j^***;**

Figure 3. Max_Distance function

Max_dif_Pvalue$(\boldsymbol{M}|\boldsymbol{z})$
Input: A set of records with standard features ***A***,
Number of rows (A) ***m,*** Normalized levenshtein distance ***Leven***;
Output: indices of the maximum difference of P-value and confidence level
$dif\_pvalue_j = \sum_{i \neq j} |p_{value}(i,j) - 0.05|$; $i \in z$& M(i,j)=0
$i^* \leftarrow \arg\_\max_j(dif\_pvalue_j)$
Return j^***;**

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: $\boldsymbol{X}$ as common name, a set of records with Name (X) $\boldsymbol{A}$, List of Common names $\boldsymbol{C}$, number of customer standard features $\boldsymbol{n}$, a set of standard features **Z**, Successful customer recognition **flag**;
Output: Sequence of questions (features) $\boldsymbol{SQ}$

1 $\boldsymbol{IF}\ Y \in C\ \boldsymbol{THEN}$
2 Break
3 **Else** $X \leftarrow Y$ **End**
4 $\boldsymbol{A} \leftarrow Records\ with\ name\ X$
5 $\boldsymbol{flag}, \boldsymbol{i} \leftarrow 0$
7 **While** flag = =0
8 $k \leftarrow n - i$
9 **If (k>2)**
10 $\boldsymbol{Q}^* \leftarrow \boldsymbol{ANOVA}(\mathbf{A})$
11 $\boldsymbol{SQ}(\boldsymbol{i}) \leftarrow \boldsymbol{Q}^*;\ \boldsymbol{Z} \leftarrow \boldsymbol{Z} - \{\boldsymbol{Q}^*\}$
13 **Else**
14 $\boldsymbol{Q}^* \leftarrow \mathbf{Max\_distance}(\boldsymbol{A}|\boldsymbol{z})$
15 $\boldsymbol{SQ}(\boldsymbol{i}) \leftarrow \boldsymbol{Q}^*$
16 **End**
17 $\boldsymbol{flag} \leftarrow \mathbf{Recognition}(\mathbf{A}, \boldsymbol{k}^*, \mathbf{p})$
18 $\boldsymbol{i} \leftarrow i + 1$
19 **End**
20 $\boldsymbol{A} \leftarrow$ Records with $\boldsymbol{SQ}(\boldsymbol{i})$ customer answer
21 **If** i==n
22 Go to **20**
23 **End**
24 Go to **9**
25 Ask sequence of other features
26 **END**
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

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

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

Null Hypothesis	Decision	P-value

$\mu_{street} = \mu_{suburb} = \mu_{Month}$	reject	2.20E-08

Table 11. Multiple comparisons result with correspondent P-Values

	Street	Suburb	Month
Street	**0**	**1,** 0.0049	**1,** 2.1404e-006
Suburb	**1,** 0.0049	**0**	**1,** 8.2137e-007
Month of Birth	**1,** 2.1404e-006	**1,** 8.2137e-007	**0**

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***.

Table 12. Associated Number of rejection for customer features in Multiple pair comparison test

Feature	Number of rejection
Street	2
Suburb	2
Month of Birth	2

Table 13. Max_dif_Pvalue results

Feature	Value
Street	*0.9951*
Suburb	*0.9951*
Month of Birth	*1.0000*

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 14. Updated dataset based on customer answer

	Street	Suburb	Month of Birth
2	Hart close	Dural	July
3	Neerabup	Doodenanning	July

***Table 15. Max_distance* function Result**

Feature	Value
Street	*0.9*
Suburb	*0.83*

SQ=(Month, Street)

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

[1] A. Faed, *An Intelligent Customer Complaint Management System with Application to the Transport and Logistics Industry*: Springer Science & Business, 2013.

[2] http://www.coveo.com/en/news-releases/Coveo-survey-shows-organizations-falling-short-in-generating-insight-from-unstructured-content.

[3] O. K. H. Morteza Saberi, “Intelligent Online Customer Recognition Framework: Dealing with Common Personal Names,” in ICIEA 2014, China, 2014.
[4] S. L. Pan, and J.-N. Lee, “Using e-CRM for a unified view of the customer,” *Communications of the ACM*, vol. 46, no. 4, pp. 95-99, 2003.
[5] O. K. Hussain, E. Chang, V. Ramakonar, and T. S. Dillon, "A Customer Relationship Management ecosystem that utilizes multiple sources and types of information conjointly." pp. 1-6.
[6] M. Saberi, O. K. Hussain, N. K. Janjua, and E.-J. Chang, "In-house crowdsourcing-based entity resolution: dealing with common names." pp. 83-88.
[7] R. A. Fisher, “Studies in Crop Variation. I. An examination of the yield of dressed grain from Broadbalk,” *The Journal of Agricultural Science*, vol. 11, no. 02, pp. 107-135, 1921.