Expert Systems With Applications 136 (2019) 42–49

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Genders prediction from indoor customer paths by Levenshtein-based fuzzy kNN

Onur Dogan[a,*], Basar Oztaysi[b]

[a] University of Bakircay, Department of Industrial Engineering, Izmir 35665, Turkey
[b] Istanbul Technical University, Department of Industrial Engineering, Istanbul 34367, Turkey

ARTICLE INFO

Article history:
Received 3 February 2019
Revised 11 June 2019
Accepted 14 June 2019
Available online 15 June 2019

Keywords:
Gender prediction
Path prediction
Fuzzy sets
Fuzzy kNN
Indoor paths
Levenshtein distances

ABSTRACT

Companies have an advantage over the competitors if they can present customized offers to customers. Demographic information of customers is critical for the companies to develop individualized systems. While current technologies make it easy to collect customer data, the main problem is that demographic data are usually incomplete. Hence, several methods are developed to predict unknown genders of customers. In this study, customer genders are predicted from their paths in a shopping mall using fuzzy sets. A fuzzy classification method based on Levenshtein distance is developed for string data that refer to the indoor customer paths. Although there are several ways to predict the gender, no study has focused on path-based gender classification. The originality of the research is to classify customer data into the gender classes using indoor paths.

1. Introduction

Smartphones have an important place in users daily life because smartphone usage has increased for the last decade. Researchers have usually preferred smartphones to collect ubiquitous and unobtrusive data (Fernandez-Llatas, Lizondo, Monton, Benedi, & Traver, 2015). Devices that have different technologies interact with smartphones via sensors. While it allows spatiotemporal data collection, the collected data do not include demographic information. Since demographic information is significant in individualized systems, the number of gender prediction studies have been increased in recent years. The demographic information such as gender, age, marital status, education level, location, and job are used to determine recommendations (Choi, Oh, Kim, & Ryu, 2016a), future prediction (Boland, Riggs, & Anderson, 2018) and marketing decisions (Lim, Choi, Akhmedov, & Chung, 2018). The gender information occupies an important place among demographic information affecting decision-making (Chong, 2013; Yeh, Wang, & Yieh, 2016)

Because many sensor data do not include gender information, researchers have focused on to predict the gender of users. Sensors collect spatiotemporal data for each user from smartphones as long as the users allow the interaction. These event logs enable to discover users paths, backgrounds, and interests (Zhong, Tan, Mo, & Yang, 2013). For a shopping mall example, female customers may visit more accessory stores than male customer. Male customers may have a higher duration in technology stores than female customers. Gender prediction is one of the most difficult problems due to unknown changes in users preferences (Moeini & Mozaffari, 2017). There are several ways to predict users gender using face (Danisman, Bilasco, & Martinet, 2015), speech (Bisio, Delfino, Lavagetto, Marchese, & Sciarrone, 2013), handwriting (Ahmed, Rasool, Afzal, & Siddiqi, 2017).

Understanding indoor customer paths is an attractive study area (Abedi, Bhaskar, Chung, & Miska, 2015; Mazimpaka & Timpf, 2016). Path analysis studies have increased popularity among researchers (Brun, Saggese, & Vento, 2014; Saini, Kumar, Roy, & Dogra, 2017). In our previous study (Dogan, Bayo-Monton, Fernandez-Llatas, & Oztaysi, 2019a), we investigated male and female customers paths to show behavioral changes. The study indicates male and female customers have different behaviors with respect to their paths which are followed in the shopping mall. In such studies, knowing the gender of customers gives advantages to understand customers better. Information in individualized behavior targeting (Hu, Zeng, Li, Niu, & Chen, 2007), web search (Weber & Castillo, 2010) and advertisement targeting (Mei, Li, Tian, Tao, & Ngo, 2018) are some of the contemporary areas that need gender information. It can provide to learn users interests. Mei et al. (2018) developed a contextual advertising platform which changes visual appearance of a webpage according to the publishers interest. Pervasive comput-

* Corresponding author.
E-mail addresses: onur.dogan@bakircay.edu.tr (O. Dogan), oztaysib@itu.edu.tr (B. Oztaysi).

https://doi.org/10.1016/j.eswa.2019.06.029
0957-4174/

ing applications may need gender information in some application. For instance, a smart device embedded into a store may send promotions according to the identified gender settings. Access control systems in shopping mall toilettes can use gender information to allow for only predefined gender entrance.

In the study, we classify indoor customer paths into male or female classes using fuzzy sets based on Levenshtein distance measure. We have chosen a fuzzy-based classification method because fuzzy methods enable more flexible and better solutions (Bhattacharya & Bhatnagar, 2012). Parameters in the algorithms require to tune to obtain high accuracy rates. Algorithms that have high number of parameters make difficult to find best results. Although some data mining techniques such as random forest that can be used for classification depend on more than one parameter, kNN has only one parameter, which is k showing the number of closest neighbors. Moreover, the fuzzy extension of kNN simplifies to reach better classification solutions.

This paper is structured as follows. Firstly, previous studies about gender classification and customer paths are examined. Also, related works with fuzzy kNN are presented. This part shows the gap in the literature. Secondly, the selected classification algorithm introduced. Thirdly, the case study is represented. We discuss the results of the algorithm in discussions and results section. Then, the conclusions of the study and the limitations and future works are given.

2. Literature review

Classification, which is a part of data analysis, is a famous method to group data into the existing classes. A classification algorithm aims to assign data into the predetermined classes with respect to the similarity. Classifying the customer paths to find the customers gender is attractive research. According to our research on previous studies, which is given following subsections, there has not been any study which classifies the indoor paths to predict the customers gender using Levenshtein-based fuzzy kNN method. Because the benefits of knowing the customers gender provides various advantages which are discussed before, the study fills a gap in the literature by improving previous researches.

2.1. Genders classification and path studies

Gender classification is one of the implementation areas of classification studies. Researchers who do not use smartphones to collect data have focused on gender classification in different ways. Face (Chen & Ross, 2011; Lu, Chen, & Jain, 2006), speech (Bisio et al., 2013), handwriting (Ahmed et al., 2017; Al Maadeed & Hassaine, 2014; Bouadjenek, Nemmour, & Chibani, 2014) and video-based gait (Hu, Wang, Zhang, & Zhang, 2011; Li, Maybank, Yan, Tao, & Xu, 2008b) are used to gender classification.

Available gender classification approaches in smartphones are performed in two ways: visual or audio. Danisman et al. (2015) developed a fuzzy inference system that uses some facial features such as inner face and mustache. Agneessens, Bisio, Lavagetto, and Marchese (2010) collect audio signals from mobile phone and describe speakers gender. Choi, Kim, Kim, Park, and Park (2016b) analyze mobile text data to be able to predict users gender. Table 1 shows the related works on gender classification with a different modality. Because some researchers check the accuracy of the proposed methodology for several datasets, the accuracy of the study may have an interval value. It indicates the smallest and largest accuracy rates for different training datasets in Success column.

Many studies use customer path for several purposes such as path-based prediction, path-based recommendation, and mining frequent patterns. Table 2 presents that studies using user paths have not focused on gender prediction. Studies on gender prediction have not used user paths, and studies that use user paths have not focused on gender prediction. Therefore, gender classification from customer user paths is the contribution of the research.

2.2. Fuzzy kNN-based studies

k-Nearest Neighbor (kNN) is one of the classification methods in data mining. It has been a hot topic both in data mining and machine learning researches. kNN algorithm was firstly proposed by Cover and Hart (1967). One disadvantage of the method is each labeled data has equal importance in the classification of the samples. However, the typicalness of the samples has also significance (Keller, Gray, & Givens, 1985). Over the years, many researchers have developed the kNN algorithm. One of the extensions of kNN includes fuzzy sets to overcome uncertainty in crisp classification. Keller et al. (1985) proposed a fuzzy version of the kNN algorithm, which adjusts the membership values between zero and one according to how correctly classified. The membership value of a sample closes to one if the sample correctly classified. They showed that the developed fuzzy kNN (FkNN) algorithm reasonably assigns the membership values and produces lower error rates. Warren and Damin (1997) applied the fuzzy kNN in two manufacturing cases to identify welds from digital images and to determine failures in face milling operations. They obtained success rates with 93.2% and 89.1% by eliminating overlapped samples.

Calculating fuzzy membership degrees for each instance to the classes needs computational cost and time which change according to the volume of data. Taneja, Gupta, Aggarwal, and Jindal (2015) modified fuzzy kNN algorithm, which is called MFZ-KNN, to reduce complexity and computational time. The developed algorithm clusters instances in preprocessing step and the fuzzy membership values are calculated in reference to cluster centers. Maillo, Luengo, García, Herrera, and Triguero (2017) used datasets including about 11 million data to run fuzzy kNN classifier on big datasets. They produced the same precision rate with the original algorithm in less time.

FkNN has been successfully applied in many areas, such as astronomy (Mohamed, 2018), banking and finance (Chen et al., 2011), biology (Shen, Yang, & Chou, 2006), biometric (Arai et al., 2010; Liew, Choo, Low, & Yusoh, 2017; Wang, Liu, & Zhang, 2015), healthcare (Chen et al., 2013), materials science (Tu, Zhao, Liu, Shen, & Yu, 2011) and so on. However, indoor customer paths have not classified with fuzzy kNN method. Since assignment a sample to one of the classes may be different from zero-one membership degree, we propose a fuzzy based k-nearest neighbor methodology that used Levenshtein distances, which is one of the edit-based distance methods, to classify customer paths in a shopping mall.

3. Methodology

In this paper, a novel method is proposed for gender prediction by using indoor locations of customers. The proposed method is based on Levenshtein distance and fuzzy kNN. Fig. 1 indicates the flowchart of the proposed method.

The proposed method starts with collection of beacon data from customers' mobile phone. Then the data is converted to store and store group data. Some data preparation steps are applied to clean the data. And then, individual paths of the customers are determined. Later, the paths are grouped into gender patterns. The distances between a gender-unknown path and all gender-known paths are calculated by using Levenshtein distance. Finally, the fuzzy nearest neighbors are determined in order to predict the gender of the customer.

Table 1
A summary of gender classification studies.

Study	Modality	Mobile phone	Success
Danisman et al. (2015)	Face	YES	93.35%
Chen and Ross (2011)	Face	NO	89–91%
Lu et al. (2006)	Face	NO	9.0–0.03[a]
Bisio et al. (2013)	Audio	NO	81.50%
Agneessens et al. (2010)	Audio	YES	55–67%
Ahmed et al. (2017)	Handwriting	NO	79–85%
Bouadjenek et al. (2014)	Handwriting	NO	59–70%
Al Maadeed and Hassaine (2014)	Handwriting	NO	98.33%
Choi et al. (2016b)	Text	YES	0.87[b]
Huang, Li, and Lin (2014)	Text	NO	65–82%
Ikeda, Hattori, Ono, Asoh, and Higashino (2013)	Text	NO	85%
Antal and Nemes (2016)	Biometric	YES	57.16–64.76%
Miguel-Hurtado, Stevenage, Bevan, and Guest (2016)	Biometric	YES	78.20%
Weiss and Lockhart (2011)	Biometric	YES	71.20%
Jain and Kanhangad (2016)	Biometric	YES	77.45%

[a] The average and standard deviation of the error rates.
[b] *F*-score.

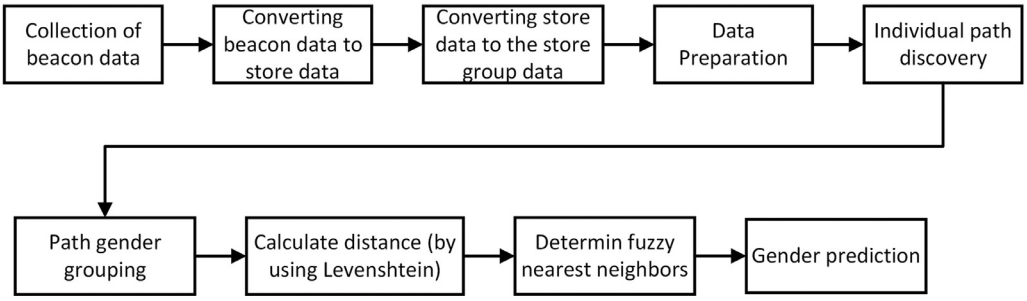

Fig. 1. Flowchart of the proposed method for gender prediction from paths.

Table 2
An overview to user path studies.

Main Focus	Study
Path-based prediction	Monreale, Pinelli, Trasarti, and Giannotti (2009); Naserian, Wang, Dahal, Wang, and Wang (2018); Ying, Lee, and Tseng (2013); Ying, Lee, Weng, and Tseng (2011)
Path-based recommendation	Brilhante, Macedo, Nardini, Perego, and Renso (2013); del Carmen Rodríguez-Hernández, Ilarri, Hermoso, and Trillo-Lado (2017); Li et al. (2008a); Xiao, Zheng, Luo, and Xie (2010); Zheng and Xie (2011); Zheng, Zhang, Ma, Xie, and Ma (2011)
Mining frequent patterns	Cao, Mamoulis, and Cheung (2005); Chen, Yuan, Qiu, and Pi (2019); Dogan, Gurcan, and Gokdemir (2019b); Lee, Chen, and Ip (2009); Shaw and Gopalan (2014)

3.1. Levenshtein distance

In the classification algorithm, Levenshtein distance, which is a specific form of sequential alignment methods, is used to compute the distance between two paths. Levenshtein distance is one of the most commonly preferred algorithms to measure differences among strings (Klomsae, Auephanwiriyakul, & Theera-Umpon, 2017). Traditional Euclidean distances are not viable for string-based data such as customer paths (DUrso & Massari, 2013). In the Levenshtein algorithm, string-based sequences are defined regarding three different transformations, which are substitution, deletion, and insertion (Levenshtein, 1966).

We point out the first path, *f*, and the second path, *s*. The mentioned three transformations are performed to a unique element of *f* and *s*. To be able to start computing the distance, the first element of string arrays is assigned to zero. Therefore, the similarity comparison matrix consists of $l_f + 1$ rows and $l_s + 1$ columns.

$f = f(0), f(1), f(2), f(l_f)$
$s = s(0), s(1), s(2), s(l_s)$

where l_f and l_s are the length of the first and second path, respectively and $f(i)$ shows ith element of the first path and $s(j)$ refers to jth element of the second path. If $Lev(f, s)$ shows the Levenshtein distance between the subsequences of the first and second path, $f(i)$ and $s(j)$, then it can be calculated as Eq. (1).

$$Lev(i,j) = min\begin{cases} Lev(i-1,j)+1 \\ Lev(i,j-1)+1 \\ Lev(i-1,j-1)+1) \end{cases} \quad (1)$$

where $Lev(i-1,j)+1$, $Lev(i,j-1)+1$, and $Lev(i-1,j-1)+1$ represent deletion, insertion and substitution operations, respectively.

3.2. Levenshtein distance based fuzzy k-nearest neighbor (L-FkNN) classification

Let $W = (x_1, x_2, x_N)$ be the dataset to be classified, a fuzzy c-partition of the dataset designates the membership of each data in each of c classes. It is denoted by the c by n matrix U, $u_{ij} = u_i(x_j)$ is the membership of x_j in the class i for $i = 1, 2, c$ and $j = 1, 2, n$. There are three basic restrictions for U to be a fuzzy c-partition.

$$\sum_{i=1}^{c} u_{ij} = 1 \quad (2)$$

$$0 < \sum_{j=1}^{n} u_{ij} < 1 \quad (3)$$

$$u_{ij} = [0, 1] \quad (4)$$

Table 3
Confusion matrix for path classification problem.

	Male	Female
Male	A	C
Female	B	D

The primary goal of the fuzzy kNN algorithm is to assign membership as a function of the data point's distance from its k-nearest neighbors and those neighbors' memberships in the possible classes (Warren & Damin, 1997). The assigned memberships are computed using Eq. (5). The membership function $u_i(x)$ assigns the values according to the inverse of the distances from the nearest neighbors. The fuzzifier m must be larger than 1. In the study, m equals to 2 so that the contribution of each neighboring point has a weight with respect to the reciprocal of its distance. $Lev(x, x_j)$ is the Levenshtein distance between the data point x_i and neighbor class x_j.

$$u_i(x) = \frac{\sum_{j=1}^{k} u_{ij}\left(\frac{1}{Lev(x,x_j)}\right)^{\frac{2}{m-1}}}{\sum_{j=1}^{k}\left(\frac{1}{Lev(x,x_j)}\right)^{\frac{2}{m-1}}} \tag{5}$$

A k-nearest neighbors rule is used to compute the membership assignments to the labeled samples. The k-nearest neighbors to each labeled sample x in class i are calculated, and then the class membership value is assigned using Eq. (6).

$$u_i(x) = \begin{cases} 0.51 + 0.49 \times \frac{n_j}{k} & if\ j = i \\ 0.49 \times \frac{n_j}{k} & if\ j \neq i \end{cases} \tag{6}$$

where n_j is the number of neighbors found which belong to the class j.

4. Path classification

Since the compared algorithms do not work well with imbalanced dataset, we balanced the training dataset to be able to compare the results. The proposed algorithm overcomes the inequality between the sizes of classes. One class may have 40 samples while other class may have 400 samples. Since the number of female customer paths is approximately four times higher than the number of male customer paths in our training dataset, the classification accuracy may be incorrect, which is called accuracy paradox. For example, let the customer path ABCD is a common path for male and female customers. In this case, the traditional algorithm can yield that a customer following the path ABCD is a female customer because it will find more female customer paths due to the higher number of female customer paths. Besides, fuzzy methods increase the correct assignment in clustering and classification problems. If a gender-unknown customer path BCBDE is seen as a female customer by the rate of 51%, traditional algorithms correctly assign it to the female customer class. They ignore being a member of the other class.

The confusion matrix is widely used in classification problems to interpret the results easily (Table 3). There are some several metrics to measure the quality of the classification algorithms (Fawcett, 2006). One of the most used quality metrics is accuracy, which defines systematic errors and measures the statistics of bias (Doğan, 2019). Recall measures the ability of an algorithm to correctly classify a female as a female (Dusetzina et al., 2014). It answers to the question: If a customer is a female, how often the algorithm yields a female. Precision, also known as the positive predictive value, represents the proportion of matched pairs classified as a female that is a female (Dusetzina et al., 2014). It answers to the question: if the test result is a female, what is the probability that the customer is female. F score is another metric of a classification problem. It is the harmonic mean of the recall and precision

$$accuracy = \frac{A + D}{A + B + C + D} \tag{7}$$

$$recall = \frac{A}{A + B} \tag{8}$$

$$precision = \frac{A}{A + C} \tag{9}$$

$$Fscore = \frac{2A}{2A + B + C} \tag{10}$$

5. Case study

5.1. Data collection and preparation

Prediction of a users gender from smartphone signals is a problem that requires a huge number of location, time and text data (Choi et al., 2016b). No study has focused on gender prediction from users path to our knowledge. The research data is collected in one of the major shopping malls in Istanbul by Bluetooth-based iBeacon devices. The shopping mall has six floors with a total of approximately 300 shops. We categorize stores into store group concerning their commodities or services. We have eleven store groups that named as Accessory (A), Catering (C), Clothing (D), Common area (E), Electronics (F), Entertainment (G), Entrance (H), Home (I), Mother&Baby (J), Personal Care (K), Supermarket (L). 3788 customer paths, 189 male customers, 704 female customers, and 2895 gender-unknown customers, are gathered. We investigate store groups instead of stores to increase understandability of the customer paths. Each store in the shopping mall grouped into a store group concerning their commodities or services.

In the preprocessing step, we ignore the paths out of 10:00:00 and 22:00:00 due to the working time of the shopping mall. If the time gap between two subsequent visits for the same customer is larger than 90 min, we assume it is a different visit belonging to the same customer. We suppose that a visit to a store must be at least 1 min. Otherwise, it means the customer data was captured while walking instead of visiting. One-location visits are ignored due to increase misclassification.

We have an imbalanced dataset. The rate of male and female paths is 26.8%, which is 189 male paths and 704 female paths. Since the incidence of the female class is dominant, incorrect classification of male paths will not affect the accuracy of the classification model, which is called accuracy paradox. It means that although the algorithm obtains a high accuracy rate, it does not reflect the real case. There are two main ways to prevent the accuracy paradox. Some additional metrics, precision, recall, and F-score, which is harmonic mean of precision and recall, can be computed. The other way is to balance the training dataset. Although the developed algorithm works with imbalanced datasets, we choose to balance dataset to be able to compare with other kinds of classification algorithms. We randomly select 26.8% of female paths and then run the algorithm. We repeated this procedure ten times.

Given the small sample size of customer path data are presented in Table 4. Each item refers to a continuous visit to a store group. For example, customer C1 visits two store groups; first store group F, then store group C. We ignore one-location visits because the prediction of them is not meaningful. The length of the paths varies from 2 to 18 locations. 99.07% of the customer paths include between two and ten store groups. C and D are two of the most visited store groups.

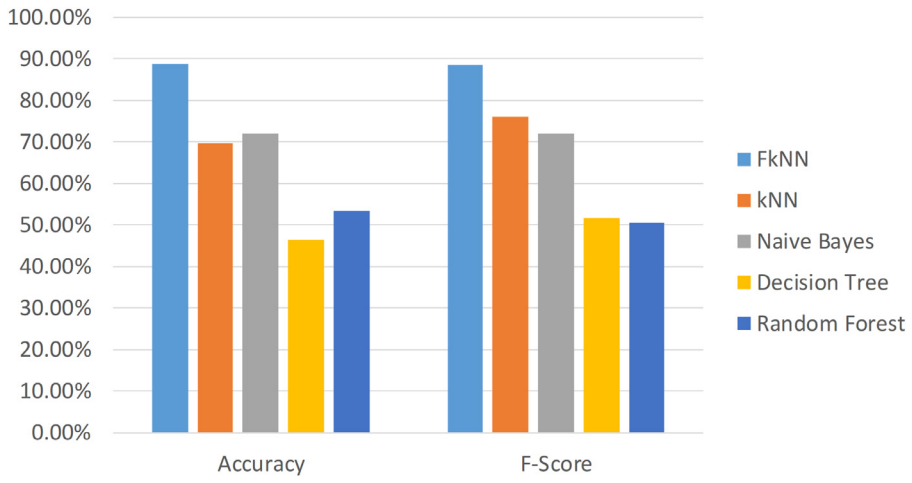

Fig. 2. Performance metrics for compared five algorithms.

Table 4
Customer path data after preprocessing step.

Customer	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
C1	C	D	C							
C2	C	D								
C3	A	D								
C4	I	K	E							
C5	I	C	K	E						
C6	J	D								
C7	L	F	G	H	J					
C8	C	K	I	E	K	G	E	K		
C9	K	G	C	K	G	H	K	G	K	H

5.2. Experimental results

Cross-validation is a way to increase the accuracy of the predictive model. We applied 10-fold cross-validation to the training dataset to avoid some problems such as overfitting (Nikoo, Kerachian, & Alizadeh, 2018). We have 893 customer paths, 189 male customers, 704 female customers. The proposed algorithm is repeated ten times by selected randomly about 189 female paths in each run. Then, we compute average accuracy and F score of the proposed algorithm, which is 87.19% and 89.11%, respectively. This means approximately every 9 out of 10 customer paths are classified correctly. As Bouadjenek et al. (2014) declare that accuracy between 59% and 70% is enough to confirm the robustness of a classification problem, we showed that our algorithm works well enough.

The Fuzzy kNN method improves in accuracy the kNN algorithm in most classification problems (Maillo, Luengo, García, Herrera, & Triguero, 2018). We compare not only kNN but also different classification methods to validate the developed method. Table 5 shows the experimental results of the training dataset.

Since the methods chosen to compare has different parameters, we give the best results obtained by changing parameters for each method. The accuracy and F score of kNN are the best among traditional methods. As expected, FkNN yields better performance result than kNN. On the other hand, kNN is the best method for precision. However, it has a low recall rate because it classified 83 customers into the female class even if they are male customer. Random forest approach gives unacceptable results for this problem. Also, decision tree algorithm results in low classification quality. In Fig. 2, which depicts the performance metrics of the compared algorithms, the x-axis shows the quality metrics and the y-axis shows the percentage of them.

Since the developed classification method FkNN predicts better the gender of customers, we apply it to the gender-unknown customer paths. Table 6 shows the results of the L-FkNN method for customer paths given in Table 4. d refers the minimum Levenshtein distance. k shows the number of the nearest neighbor. In the study, we assume $k = 3$. However, in some situations, k may have larger than three because there may have many data points with the equal minimum distances in the labeled dataset. Therefore, if the number of the nearest neighbor is less than predefined k value, the L-FkNN algorithm computes the second minimum distance and defines the data points that have those distance. For example, the minimum distance is one for the path IKE. There is only two labeled paths, which is less than three, in the training dataset. So, the second minimum distance, which is two, is calculated. After that, the total number of the nearest neighbor exceeds the predefined k value. $u_m(x)$ and $u_w(x)$ show the paths membership values belonging to the male and female class, respectively. Path Classificaiton column presents the predicted class.

As a result, the L-FkNN algorithm classifies 147 male and 2748 female paths. To obtain the gender, we need some additional data processing steps. Since one customer may visit the shopping mall more than one time, we turn into path classification to gender prediction (Fig. 3). The customer 43376598 behaves like a male in 2 visits and like a female in 4 visits. Therefore, we decide the customer is a female. Finally, genders are predicted from the path

Table 5
The performance metrics of the selected methods for the training dataset.

	L-FkNN		kNN		Naive Bayes		Decision tree		Random forest	
	W	M	W	M	W	M	W	M	W	M
Female (W)	176	21	186	11	138	59	111	86	93	104
Male (M)	22	167	83	106	49	140	68	121	113	76
Accuracy:	88.86%		75.65%		72.02%		60.10%		43.78%	
Recall:	88.89%		69.14%		73.80%		62.01%		45.15%	
Precision:	89.34%		94.42%		70.05%		56.35%		47.21%	
F score:	89.11%		79.83%		71.88%		59.04%		46.15%	

Table 6
A part of the developed L-FkNN algorithm result.

Customer	Path	d	k	$u_m(x)$	$u_w(x)$	Path classification
C1	CDC	0	3	0	1	Female
C2	CD	0	19	0.129	0.871	Female
C3	AD	0	19	0	1	Female
C4	IKE	1, 2	2, 90	0.883	0.117	Male
C5	ICKE	2	18	0.136	0.864	Female
C6	JD	0, 1	2, 110	0.088	0.913	Female
C7	LFGHJ	1	6	0.755	0.245	Male
C8	CKIEKGEK	5, 6	2, 51	0.12	0.88	Female
C9	KGCKGHKGKH	7	8	0.061	0.939	Female

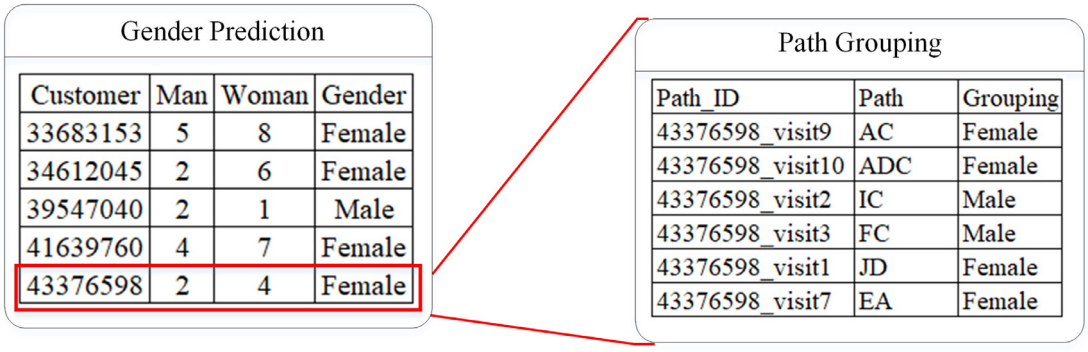

Fig. 3. Gender prediction obtained from path classification.

prediction as 99 male customers and 2124 female customers. 22 customers cannot be classified due to the equal number of path prediction.

6. Conclusions and discussions

Gender information, one of the demographic information, is critical for companies to develop individualized systems such as promotions and customer-based discounts. However, data collected by smartphones, which are very common in human life, does not include gender information. In the literature, many studies have predicted customer gender, no one has used the paths. In this study, we used ibeacon signal data provided by a beacon network implemented in an shopping mall. Due to privacy regulations, the beacon network only stores indoor location data. Therefore, the input data for our prediction problem do not include any other demographic attributes. However, some of the customer genders are known because they went to the toilette male or toilette female. We used the gender-known path as a training set. Then, we tested our method called Levenshtein based fuzzy k-nearest neighbor (L-FkNN) and some other classification methods, which are k-nearest neighbor, naive bayes, decision tree, and random forest. We validated L-FkNN yields higher classification quality that is measured by accuracy (87.19%) and F score (89.11%). Then, we applied the L-FkNN to the gender-unknown customer paths. The developed algorithm L-FkNN predicts 147 male and 2748 female paths. Because customers may have more than one visit to the shopping mall, we used the path classification to obtain gender prediction. As a result, we found there are 99 male and 2124 female customers among gender-unknown paths. 22 customer paths have an equal number of path prediction. Therefore, we cannot classify them.

Although the study gives useful results, it has some limitations. The L-FkNN classification method works for one-dimensional data, in here locations. However, the gender of the customers may depend on the spent time in locations. The developed method may result in better accuracy and F score metrics by considering the duration. In this study, we used Levenshtein distance measure. Other kinds of string distance measures such as Jaccard, Needleman-wunch, and Smith-waterman can be used to look for better results. We hope this work can serve as a starting point for providing advanced results. Fuzzy theory has many advantages. For example, it can solve problems with imprecise and incomplete data. It is readily customizable in natural language terms. On the other hand, fuzzy based studies require more run time. The L-FkNN algorithm results in an equal number of paths for male and female classes while classifying gender from path prediction like explained before. This causes to ignore some data to classify at the end of the algorithm. We grouped stores into store groups concerning their commodities or services according to our personal experience. They could be grouped more objectively, and different classification results could be obtained.

Declaration of Competing Interest

There is no conflict of interest.

References

Abedi, N., Bhaskar, A., Chung, E., & Miska, M. (2015). Assessment of antenna characteristic effects on pedestrian and cyclists travel-time estimation based on bluetooth and wifi mac addresses. *Transportation Research Part C: Emerging Technologies, 60*, 124–141.

Agneessens, A., Bisio, I., Lavagetto, F., & Marchese, M. (2010). Design and implementation of smartphone applications for speaker count and gender recognition. In *The internet of things* (pp. 187–194). Springer.

Ahmed, M., Rasool, A. G., Afzal, H., & Siddiqi, I. (2017). Improving handwriting based gender classification using ensemble classifiers. *Expert Systems with Applications, 85*, 158–168.

Al Maadeed, S., & Hassaine, A. (2014). Automatic prediction of age, gender, and nationality in offline handwriting. *EURASIP Journal on Image and Video Processing, 2014*(1), 10.

Antal, M., & Nemes, G. (2016). Gender recognition from mobile biometric data. In *Applied computational intelligence and informatics (SACI), 2016 ieee 11th international symposium on* (pp. 243–248). IEEE.

Arai, Y., Lien, N. T. H., Hiroyuki Satoh, K. I., Hayashi, T. F. D., & Hirota, K. (2010). Fuzzy few-nearest neighbor method with a few samples for personal authentication. *Journal of Advanced Computational Intelligence and Intelligent Informatics, 14*(2), 167–178. doi:10.20965/jaciii.2010.p0167.

Bhattacharya, S., & Bhatnagar, V. (2012). Fuzzy data mining: A literature survey and classification framework. *International Journal of Networking and Virtual Organisations, 11*(3–4), 382–408.

Bisio, I., Delfino, A., Lavagetto, F., Marchese, M., & Sciarrone, A. (2013). Gender-driven emotion recognition through speech signals for ambient intelligence applications.. *IEEE Transactionson Emerging Topics in Computing, 1*(2), 244–257.

Boland, J., Riggs, K. J., & Anderson, R. J. (2018). A brighter future: The effect of positive episodic simulation on future predictions in non-depressed, moderately dysphoric & highly dysphoric individuals. *Behaviour Research and Therapy, 100*, 7–16.

Bouadjenek, N., Nemmour, H., & Chibani, Y. (2014). Local descriptors to improve off-line handwriting-based gender prediction. In *Soft computing and pattern recognition (SOCPAR), 2014 6th international conference of* (pp. 43–47). IEEE.

Brilhante, I., Macedo, J. A., Nardini, F. M., Perego, R., & Renso, C. (2013). Where shall we go today?: Planning touristic tours with tripbuilder. In *Proceedings of the 22nd ACM international conference on information & knowledge management* (pp. 757–762). ACM.

Brun, L., Saggese, A., & Vento, M. (2014). Dynamic scene understanding for behavior analysis based on string kernels. *IEEE Transactions on Circuits and Systems for Video Technology, 24*(10), 1669–1681.

Cao, H., Mamoulis, N., & Cheung, D. W. (2005). Mining frequent spatio-temporal sequential patterns. In *Data mining, fifth IEEE international conference on* (pp. 8–pp). IEEE.

del Carmen Rodríguez-Hernández, M., Ilarri, S., Hermoso, R., & Trillo-Lado, R. (2017). Towards trajectory-based recommendations in museums: Evaluation of strategies using mixed synthetic and real data. *Procedia Computer Science, 113*, 234–239.

Chen, C., & Ross, A. (2011). Evaluation of gender classification methods on thermal and near-infrared face images. In *Biometrics (IJCB), 2011 international joint conference on* (pp. 1–8). IEEE.

Chen, H.-L., Huang, C.-C., Yu, X.-G., Xu, X., Sun, X., Wang, G., et al. (2013). An efficient diagnosis system for detection of Parkinsons disease using fuzzy k-nearest neighbor approach. *Expert Systems with Applications, 40*(1), 263–271.

Chen, H.-L., Liu, D.-Y., Yang, B., Liu, J., Wang, G., & Wang, S.-J. (2011). An adaptive fuzzy k-nearest neighbor method based on parallel particle swarm optimization for bankruptcy prediction. In *Pacific-asia conference on knowledge discovery and data mining* (pp. 249–264). Springer.

Chen, Y., Yuan, P., Qiu, M., & Pi, D. (2019). An indoor trajectory frequent pattern mining algorithm based on vague grid sequence. *Expert Systems with Applications, 118*, 614–624.

Choi, I. Y., Oh, M. G., Kim, J. K., & Ryu, Y. U. (2016a). Collaborative filtering with facial expressions for online video recommendation. *International Journal of Information Management, 36*(3), 397–402.

Choi, Y., Kim, Y., Kim, S., Park, K., & Park, J. (2016b). An on-device gender prediction method for mobile users using representative wordsets. *Expert Systems with Applications, 64*, 423–433.

Chong, A. Y.-L. (2013). Predicting m-commerce adoption determinants: A neural network approach. *Expert Systems with Applications, 40*(2), 523–530.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21–27.

Danisman, T., Bilasco, I. M., & Martinet, J. (2015). Boosting gender recognition performance with a fuzzy inference system. *Expert Systems with Applications, 42*(5), 2772–2784.

Doğan, O. (2019). Data linkage methods for big data management in industry 4.0. In *Optimizing big data management and industrial systems with intelligent techniques* (pp. 108–127). IGI Global.

Dogan, O., Bayo-Monton, J.-L., Fernandez-Llatas, C., & Oztaysi, B. (2019a). Analyzing of gender behaviors from paths using process mining: A shopping mall application. *Sensors, 19*(3), 557.

Dogan, O., Gurcan, O. F., Oztaysi, B., & Gokdemir, U. (2019b). Analysis of frequent visitor patterns in a shopping mall. In *Industrial engineering in the big data era* (pp. 217–227). Springer.

Dusetzina, S. B., Tyree, S., Meyer, A.-M., Meyer, A., Green, L., & Carpenter, W. R. (2014). *Linking data for health services research: A framework and instructional guide*. (US), Rockville (MD): Agency for Healthcare Research and Quality.

DUrso, P., & Massari, R. (2013). Fuzzy clustering of human activity patterns. *Fuzzy Sets and Systems, 215*, 29–54.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters, 27*(8), 861–874.

Fernandez-Llatas, C., Lizondo, A., Monton, E., Benedi, J.-M., & Traver, V. (2015). Process mining methodology for health process tracking using real-time indoor location systems. *Sensors, 15*(12), 29821–29840.

Hu, J., Zeng, H.-J., Li, H., Niu, C., & Chen, Z. (2007). Demographic prediction based on user's browsing behavior. In *Proceedings of the 16th international conference on world wide web* (pp. 151–160). ACM.

Hu, M., Wang, Y., Zhang, Z., & Zhang, D. (2011). Gait-based gender classification using mixed conditional random field. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 41*(5), 1429–1439.

Huang, F., Li, C., & Lin, L. (2014). Identifying gender of microblog users based on message mining. In *International conference on web-age information management* (pp. 488–493). Springer.

Ikeda, K., Hattori, G., Ono, C., Asoh, H., & Higashino, T. (2013). Twitter user profiling based on text and community mining for market analysis. *Knowledge-Based Systems, 51*, 35–47.

Jain, A., & Kanhangad, V. (2016). Investigating gender recognition in smart-phones using accelerometer and gyroscope sensor readings. In *2016 international conference on computational techniques in information and communication technologies (ICCTICT)* (pp. 597–602).

Keller, J. M., Gray, M. R., & Givens, J. A. (1985). A fuzzy k-nearest neighbor algorithm. *IEEE Transactions on Systems, Man, and Cybernetics*, (4), 580–585.

Klomsae, A., Auephanwiriyakul, S., & Theera-Umpon, N. (2017). A string grammar fuzzy-possibilistic c-medians. *Applied Soft Computing, 57*, 684–695.

Lee, A. J., Chen, Y.-A., & Ip, W.-C. (2009). Mining frequent trajectory patterns in spatial–temporal databases. *Information Sciences, 179*(13), 2218–2231.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10*(8), 707–710.

Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., & Ma, W.-Y. (2008a). Mining user similarity based on location history. In *Proceedings of the 16th ACM sigspatial international conference on advances in geographic information systems* (p. 34). ACM.

Li, X., Maybank, S. J., Yan, S., Tao, D., & Xu, D. (2008b). Gait components and their application to gender recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 38*(2), 145–155.

Liew, S.-H., Choo, Y.-H., Low, Y. F., & Yusoh, Z. I. M. (2017). Eeg-based biometric authentication modelling using incremental fuzzy-rough nearest neighbour technique. *IET Biometrics, 7*(2), 145–152.

Lim, H., Choi, J.-G., Akhmedov, A., & Chung, J. (2018). Predicting future trends of media elements in hotel marketing by using change propensity analysis. *International Journal of Hospitality Management*.

Lu, X., Chen, H., & Jain, A. K. (2006). Multimodal facial gender and ethnicity identification. In *International conference on biometrics* (pp. 554–561). Springer.

Maillo, J., Luengo, J., García, S., Herrera, F., & Triguero, I. (2017). Exact fuzzy k-nearest neighbor classification for big datasets. In *Fuzzy systems (fuzz-IEEE), 2017 IEEEinternational conference on* (pp. 1–6). IEEE.

Maillo, J., Luengo, J., García, S., Herrera, F., & Triguero, I. (2018). A preliminary study on hybrid spill-tree fuzzy k-nearest neighbors for big data classification. In *2018 IEEE international conference on fuzzy systems (fuzz-IEEE)* (pp. 1–8). IEEE.

Mazimpaka, J. D., & Timpf, S. (2016). Trajectory data mining: A review of methods and applications. *Journal of Spatial Information Science, 2016*(13), 61–99.

Mei, T., Li, L., Tian, X., Tao, D., & Ngo, C.-W. (2018). Pagesense: Toward stylewise contextual advertising via visual analysis of web pages. *IEEE Transactions on Circuits and Systems for Video Technology, 28*(1), 254–266.

Miguel-Hurtado, O., Stevenage, S. V., Bevan, C., & Guest, R. (2016). Predicting sex as a soft-biometrics from device interaction swipe gestures. *Pattern Recognition Letters, 79*, 44–51.

Moeini, H., & Mozaffari, S. (2017). Gender dictionary learning for gender classification. *Journal of Visual Communication and Image Representation, 42*, 1–13.

Mohamed, T. M. (2018). Pulsar selection using fuzzy knn classifier. *Future Computing and Informatics Journal, 3*(1), 1–6.

Monreale, A., Pinelli, F., Trasarti, R., & Giannotti, F. (2009). Wherenext: A location predictor on trajectory pattern mining. In *Proceedings of the 15th ACM sigkdd international conference on knowledge discovery and data mining* (pp. 637–646). ACM.

Naserian, E., Wang, X., Dahal, K., Wang, Z., & Wang, Z. (2018). Personalized location prediction for group travellers from spatial–temporal trajectories. *Future Generation Computer Systems, 83*, 278–292.

Nikoo, M. R., Kerachian, R., & Alizadeh, M. R. (2018). A fuzzy knn-based model for significant wave height prediction in large lakes. *Oceanologia, 60*(2), 153–168.

Saini, R., Kumar, P., Roy, P. P., & Dogra, D. P. (2017). An efficient approach for trajectory classification using fcm and svm. In *IEEE region 10 symposium (TENSYMP), 2017* (pp. 1–4). IEEE.

Shaw, A. A., & Gopalan, N. (2014). Finding frequent trajectories by clustering and sequential pattern mining. *Journal of Traffic and Transportation Engineering (English Edition), 1*(6), 393–403.

Shen, H.-B., Yang, J., & Chou, K.-C. (2006). Fuzzy knn for predicting membrane protein types from pseudo-amino acid composition. *Journal of Theoretical Biology, 240*(1), 9–13.

Taneja, S., Gupta, C., Aggarwal, S., & Jindal, V. (2015). Mfz-knna modified fuzzy based k nearest neighbor algorithm. In *Cognitive computing and information processing (CCIP), 2015 international conference on* (pp. 1–5). IEEE.

Tu, D. C., Zhao, J. H., Liu, M. H., Shen, J., & Yu, F. (2011). Preliminary study on quantification of duck color based on fuzzy k–nearest neighbor method. In *Applied mechanics and materials* (pp. 210–215). Trans Tech Publ. 39

Wang, X.-H., Liu, A., & Zhang, S.-Q. (2015). New facial expression recognition based on fsvm and knn. *Optik-International Journal for Light and Electron Optics, 126*(21), 3132–3134.

Warren, L. T., & Damin, L. (1997). Two manufacturing applications of the fuzzy k-nn algorithm. *Fuzzy Sets and Systems, 92*(3), 289–303.

Weber, I., & Castillo, C. (2010). The demographics of web search. In *Proceedings of the 33rd international ACMSIGIR conference on research and development in information retrieval* (pp. 523–530). ACM.

Weiss, G. M., & Lockhart, J. W. (2011). Identifying user traits by mining smart phone accelerometer data. In *Proceedings of the fifth international workshop on knowledge discovery from sensor data* (pp. 61–69). ACM.

Xiao, X., Zheng, Y., Luo, Q., & Xie, X. (2010). Finding similar users using category-based location history. In *Proceedings of the 18th sigspatial international conference on advances in geographic information systems* (pp. 442–445). ACM.

Yeh, C.-H., Wang, Y.-S., & Yieh, K. (2016). Predicting smartphone brand loyalty: Consumer value and consumer-brand identification perspectives. *International Journal of Information Management, 36*(3), 245–257.

Ying, J. J.-C., Lee, W.-C., & Tseng, V. S. (2013). Mining geographic-temporal-semantic patterns in trajectories for location prediction. *ACM Transactions on Intelligent Systems and Technology (TIST), 5*(1), 2.

Ying, J. J.-C., Lee, W.-C., Weng, T.-C., & Tseng, V. S. (2011). Semantic trajectory mining for location prediction. In *Proceedings of the 19th ACM sigspatial international conference on advances in geographic information systems* (pp. 34–43). ACM.

Zheng, Y., & Xie, X. (2011). Learning travel recommendations from user-generated gps traces. *ACM Transactions on Intelligent Systems and Technology (TIST), 2*(1), 2.

Zheng, Y., Zhang, L., Ma, Z., Xie, X., & Ma, W.-Y. (2011). Recommending friends and locations based on individual location history. *ACM Transactions on the Web (TWEB), 5*(1), 5.

Zhong, E., Tan, B., Mo, K., & Yang, Q. (2013). User demographics prediction based on mobile data. *Pervasive and Mobile Computing, 9*(6), 823–837.