AP Selection for Indoor Localization Based on Neighborhood Rough Sets

Zhu Yu-jia
School of Electronic Engineering
Beijing University of Posts and Telecommunications
Beijing, China
zhuyjbupt@gmail.com

Deng Zhong-liang[1,2]
[1]Key Laboratory of Information Photonics and Optical Communications
[2]School of Electronic Engineering
Beijing University of Posts and Telecommunications
Beijing, China
dengzhl@bupt.edu.cn

Abstract—**In this paper, a new AP (access points) selection method is proposed based on neighborhood rough sets (NRS) for indoor localization. Due to the existence of virtual APs in large scale buildings, received signal strength (RSS) may produce similar measurements, leading to biased estimates and redundant computations. Our method utilize neighborhood relations to transform the radio map into an extended rough set model such that offers a more discriminative way to select a small subset of available APs. Experimental results show that our method can be used to further improve the performance of the original AP selection methods.**

Keywords- Neighborhood Rough Sets (NRS); access point (AP) selection; indoor positioning; Wireless Local Area Network(WLAN)

I. INTRODUCTION

Recently, Wireless Local Area Network (WLAN) has gained significant interest on how to design accurate and low cost sensor localization system for many personal and commercial applications [1][2]. There are three major location metrics in WLAN: the time of arrival, angle of arrival, and Received Signal Strength (RSS). Among these algorithms, RSS-based methods have been extensively studied as an inexpensive solution for indoor positioning system [3][4]. Compared with other algorithms, RSS can be easily obtained by WLAN integrated mobile device without any additional hardware modification.

The RSS information for indoor location applications has been processed in two different ways. One method is the physical radio propagation model. Because of the complexity of the radio propagation in the indoor scenario [5], the model cannot be precisely described. The other is called fingerprint method. A database of RSS measurements (also called radio map) made at a set of known locations can be initially assembled. This is done at an offline stage, and the resultant database is used as the training set for a statistical learning model [5][6]. Then, at an online stage, the learning model is used to estimate a location from a given new set of RSS values. This technique, known as fingerprinting, generally overcomes several limitations of the propagation model-based approaches, especially in complex scenarios.

The key technical challenge in fingerprint method localization is how to map the measured RSS vector received from access points (APs) to a spatial position in 2D Cartesian coordinates [7], which can be described as the mapping rule $H(\cdot): R^n \to R^2$. In order to model this mapping, lots of pattern matching methods are proposed. One simple solution is the K-nearest neighbor (KNN) method [8], which estimates the position by computer the k closest neighbors with smallest Euclidean distance using the offline collected RSS database. Statistical method is proposed to analyze the probability of each potential position such as maximum likelihood (ML) algorithm [9]. Kernel method [10] is another solution that mapping the original RSS vector into a kernel feature space for better estimation.

Except for the precision, computational complexity and the storage capacity also need to be jointly considered, especially on smart phones. Researchers have studied lots of ways to reduce the computing cost. Such as spatial filtering [4] and offline clustering [9][11]. However due to the wide deployment of APs, the dimension of the measurement vector is generally much greater than the minimum needed for positioning.

There are several ways to make AP selection, Youssef et.al bring up the Maxmean method [9]. It sort APs in descending order and select the k strongest APs. Chen et.al's Info Gain method [12] offers a novel selection strategy based on the discriminant power of each AP using the information entropy. Kushki et.al [4] demonstrate minimizing correlation between selected APs, the selection strategies are designed to mirror the properties of the distance calculation schemes under investigation. Fang et.al [13] make location fingerprinting in a decorrelated space making use of PCA, aiming to project the data into a lower dimensional space. It not only can reduce the computational complexity, but also reserving all the APs' information.

978-1-4673-1881-5/12/$31.00 ©2012 IEEE

Due to real situation of building interiors, we have noted the existence of the virtual APs (VAP). We all know different AP has different MAC address. That is the reason why we use MAC addresses to identify different APs. But if we scan the WIFI signal inside a building, we may find several signal strength change almost the same from different MAC address. It seems these APs placed overlap in the same location. The reason why we may meet such "strange" APs lies in the fact that AP is not mainly used for indoor positioning，but for communication. A lot of large-scale constructions use enterprise-class APs to build wireless lan. This kind of AP generally supports more than one VAP. Each VAP specifies its own SSID and MAC. So the identity of access for different SSID can be configured for different types of wireless users. In such situation, if we still think that different MAC stands for different placed AP, and take it as a kind of feature, apparently directly using any kind of AP selection methods above seems kind of unreasonable.

Considering such distinct APs may produce similar measurements, leading to biased estimates and redundant computations. Using AP selection techniques to select a subset of available APs for positioning is very practical.

The rest of this paper is organized as follows. Section II details our method based on Neighborhood Rough Sets (NRS). Experiments are presented to show superiority of our method over other state-of-the-art methods in Section III. Finally, Section IV gives a conclusion.

II. PROPOSED METHOD

A. NRS

Rough sets theory, introduced by Pawlak in 1982, is a powerful mathematical tool for dealing with inconsistent information in decision situations [14]. Nowadays, many rough sets based approaches have been successfully applied in feature selection, rule learning, and classifier design. The idea of rough set theory is based on the indiscernibility relation, and any set of all indiscernible objects is called an elementary set, which forms a basic granule of knowledge in a universe. However, the classical rough set model employs equivalence relations to partition the universe and generate mutually exclusive equivalence classes as elemental concepts. This is just applicable to data with nominal attributes whereas in many cases we have to deal with numerical attributes since practical data are always numerical [15]. Therefore, neighborhood relations have been proposed as an extended rough set model in the information system and have successfully applied in many fields [16].

The radio map contains a set of RSS measurements for n locations in the environment, designated as reference points (RF). Denote this map as

$$\mathrm{S}=\left\{\left\{\mathrm{p}_1,\mathrm{l}_1,\mathrm{r}_{11},\ \ldots,\ \mathrm{r}_{1m}\right\},\ \ldots,\ \left\{\mathrm{p}_\mathrm{n},\mathrm{l}_\mathrm{n},\mathrm{r}_{\mathrm{n}1},\ \ldots,\ \mathrm{r}_{nm}\right\}\right\}$$

where, r_{ij} is a vector of readings from m MAC addresses at RP p_i (p_i is a Cartesian coordinate) with label l_i. Each MAC address is a feature. Let $M=\{mac_1,mac_2,...,mac_m\}$ denotes m features. We collect RSS measurements for cc times at each RF.

Given an information system for classification learning $NDT=\langle S_{RP},M\cup D,V,f\rangle$

where $S_{RP}=\begin{Bmatrix}\{\mathrm{r}_{11},\cdots,\mathrm{r}_{1m}\}^1,\cdots,\{\mathrm{r}_{11},\cdots,\mathrm{r}_{1m}\}^{cc}\\ \{\mathrm{r}_{21},\cdots,\mathrm{r}_{2m}\}^1,\cdots,\{\mathrm{r}_{21},\cdots,\mathrm{r}_{2m}\}^{cc}\\ \cdots\\ \{\mathrm{r}_{\mathrm{n}1},\cdots,\mathrm{r}_{nm}\}^1,\cdots,\{\mathrm{r}_{\mathrm{n}1},\cdots,\mathrm{r}_{nm}\}^{cc}\end{Bmatrix}$ is a nonempty sample set called sample space, $M=\{mac_1,mac_2,...,mac_m\}$ is a nonempty set of features also called condition attributes to characterize the samples, $D=\begin{Bmatrix}l_1,\cdots,l_1\\ l_2,\cdots,l_2\\ \cdots\\ \underbrace{l_\mathrm{n},\cdots,l_\mathrm{n}}_{cc}\end{Bmatrix}$ is a set of output variable called decision attribute (class labels of RF samples), V_a is a value domain of attribute $a\in M\cup D$, f is an information function $f:S_{RP}\times(M\cup D)\to V$, $V=\cup_{a\in M\cup D}V_a$, a reduction is a minimal set of attributes $B\subseteq M$.

Given for all $s_i\in S_{RP}$ and $B\subseteq M$, The neighborhood $\delta_\mathrm{B}(s_i)$ of in the feature space B is defined as

$$\delta_\mathrm{B}(s_i)=\{s_j\in S_{RP}\mid\Delta_B(s_i,s_j)\le\delta\}\qquad(1)$$

δ is the threshold and $\Delta_B(s_i,s_j)$ is the metric function in subspace B. There are three common metric functions that are widely used. Let s_1 and s_2 be two samples in m dimensional space $M=\{mac_1,mac_2,...,mac_m\}$. $f(s,mac_i)$ denotes the value s_{is} of mac_i in the sample s. Then Minkowsky distance is defined as

$$\Delta_P(s_1,s_2)=(\sum_{j=1}^{m}|f(s_1,mac_j)-f(s_2,mac_j)|)^{1/P}\qquad(2)$$

where (1) if p = 1, it is called Manhattan distance Δ_1; (2)if p = 2, it is called Euclidean distance Δ_2; (3) if p = ∞,it is called Chebychev distance. Here, we use the Manhattan distance.

Given a neighborhood decision table NDT, $S_1,S_2,...,S_n$ are the sample subsets with decisions 1 to n, $\delta_\mathrm{B}(s_i)$ is the neighborhood information granules including s_i, and is generated by feature subset $B\subseteq M$， then the lower and upper approximations of the decision D with respect to MAC addresses subset B are, respectively, defined as

$$L_B(D)=\bigcup_{i=1}^{n}L_B(S_i)$$

$$U_B(D)=\bigcup_{i=1}^{n}U_B(S_i)\qquad(3)$$

where $L_B(S)=\{s_i\mid\delta_B(x_i)\subseteq S,x_i\in S_{RP}\}$ is the lower approximations of the sample subset S with respect to MAC addresses subset B, and is also called positive region denoted by $Pos_B(D)$ which is the sample set that can be classified into one of the classes without uncertainty with the MAC addresses subset

B .

$U_B(S)=\{s_i \mid \delta_B(x_i)\cap S\neq\phi, x_i\in S_{RP}\}$ denotes the upper approximations, obviously $U_B(S)=S_{RP}$ .The decision boundary region of D to B is defined as

$$BN_B(D)=U_B(D)-L_B(D)\ . \tag{4}$$

The greater the positive region is, the smaller the boundary region will be, and the stronger the characterizing power of the condition attributes will be. So we use the dependency degree of D to B to characterize the power of the selected MAC addresses subsets, which is defined as the ratio of consistent objects

$$\gamma_B(D)=\frac{Card(Pos_B(D))}{Card(S_{RP})}, \tag{5}$$

where $Card(S_{RP})$ and $Card(Pos_B(D))$ denotes the cardinal number of sample set S_{RP} and $Pos_B(D)$, respectively. Where $\gamma_B(D)$ reflects the ability of B to approximate D. Obviously, $0\le\gamma_B(D)\le 1$.

B. AP selection based on NRS

Initially, all the features in $M=\{mac_1,mac_2,...,mac_m\}$ are on the candidate list CL . RED is an empty pool to contain the selected features. $SMP^0=S_{RP}$ is a sample list to be discriminated by RED . At first RED^0 is empty. All we want to do is to select the most discriminational MAC addresses subsets from M. The significance of a feature mac_i as

$$SNFC(mac_i,B,D)=\gamma_{B\cup mac_i}(D)-\gamma_B(D)\ ,\ mac_i\notin B$$

We say feature mac_i is superfluous in B with respect to D if $SNFC(mac_i,B,D)=0$, otherwise mac_i is indispensable. We say B is dependent if $\forall mac_i\in B$, otherwise mac_i is indispensable.

It is a combinational optimization problem to find all of the reducts. There are $2^m-1(=C_m^1+...+C_m^i+...+C_m^m)$, $i\in(1,m)$ combinations of feature sub sets. It is not practical to search all of the reducts in 2^m-1 combinations. In practice, we usually just require one of the reducts to train a classifier. So we can use greedy forward search algorithm beginning with the feature which has the highest value of RSS.

First we should go through all the features in M and find the $mac_i\in M$ which has the highest value of $\gamma(D)$. $RED^1=\{mac_i\}$, $CL=\{mac_1,mac_2,...,mac_{i-1},mac_{i+1},...,mac_m\}$.

We get $SNFC(mac_i,RED^1,D)=\gamma_{RED^1}(D)-\gamma_{RED^0}(D)$ and POS^1 (samples in SMP which have the same decision attribute within δ). Cause the samples in POS_1 can be correctly discriminated by feature mac_i , we just need to discriminate the samples which cannot be correctly discriminated, that is the set of $SMP^1=SMP^0-POS^1$. The next iteration, selected feature mac_j will be the feature has the maximum $\gamma_{RED^2}(D)$, $(RED^2=RED^1\cup mac_j)$ with the constrain:

$$SNFC(mac_j,RED^2,D)=\gamma_{RED^2}(D)-\gamma_{RED^1}(D)>0\ .$$

Now the element of RED has 2 MAC addresses, $CL=\{mac_1,...,mac_{i-1},mac_{i+1},mac_{j-1},mac_{j+1},...,mac_m\}$, and $SMP^2=SMP^1-POS^2$ .The search process continues following the above iteration until meeting the stop criteria. We set the stop criteria as the feature has the maximum $\gamma_{RED^k}(D)$ with $SNFC(mac_k,RED^k,D)=\gamma_{RED^k}(D)-\gamma_{RED^{k-1}}(D)=0$ or $SMP=\phi$.

Then we can use Maxmean , PCA or other methods to do further selection based on RED , the set of selected features.

Algorithm 1 summarizes the AP selection procedure using NRS method for indoor localization.

Algorithm 1 AP selection based on NRS

Input $\langle S_{RP},M,D\rangle$, δ , γ_thd // δ is the threshold to control the size of the neighborhood, γ_thd is the threshold of stop.

Output RED is the pool to contain the selected MAC address subsets.

Step 1: Normalize collected RSS signal to obtain the sample set S_{RP} .

$\phi\rightarrow RED$; $M\rightarrow CL$; $S_{RP}\rightarrow SMP$; $\phi\rightarrow POS^0(D)$;

Step 2: iter = 1; //The times of iteration.

Step 3: While $SMP\neq\phi$

For each $mac_i\in CL$

For each $s_j\in SMP$

Calculate $POS(i)=POS(i)\cup s_j$;

End

End

End

Step 4: $\mathrm{M\_Inx}=\mathrm{Index}\left(\max\left(\mathrm{Card}(\mathrm{POS})\right)\right)$;

$RED\cup mac_{\mathrm{M\_Inx}}\rightarrow RED$; //Adding MAC address as elements of RED

$CL-mac_{\mathrm{M\_Inx}}\rightarrow CL$;

$\gamma_{mac_{\mathrm{M\_Inx}}}(D)=\frac{Card(POS^2)}{Card(SMP)}$

$SMP-POS^2\rightarrow SMP$;

Step 5: If ($\gamma_{mac_{\mathrm{M\_Inx}}}(D)<\gamma\_thd$) or (iter > Depth)

Break;

Else

iter = iter + 1;

Go to step 3;

End

Step 6: Use the other methods to select MAC addresses based on RED .

III. EXPERIMENTS

A. Experiments Setup

RSS data were recorded in a realistic WLAN indoor environment, as shown in Fig.1. The dimension of environment is about 215m × 142m, but we just test a subset area of about

48m×48m. Every location in the environment was covered by lots of IEEE 802.11b/g APs. A total of 125 APs (different MAC addresses) were detected throughout the test area (some areas we could only detect 76 different MAC addresses). A PDA (HTC HD2 with Windows Mobile 6.0) was used to measure WLAN signal strength value. The RSS data were collected on the device by using the open source library OpenNetCF, which provides access to MAC address and RSS values of WLAN APs. 31 reference points (RFs) were selected. Due to the area is rather large, for reducing the collection load, we choose 6m as the grid spacing. We collected 100 RSS samples per RF for training and 50 random positions to test the method at a rate of 1 sample/sec (total 3,100 training samples and 50 test samples). -100dbm were set for RPs cannot detect the specified AP. We adopted KNN classifier to evaluate the classification performance of the selected MAC addresses subsets. The distance error is adopted as the performance metric which is the Euclidean distance between the estimated and the true position.

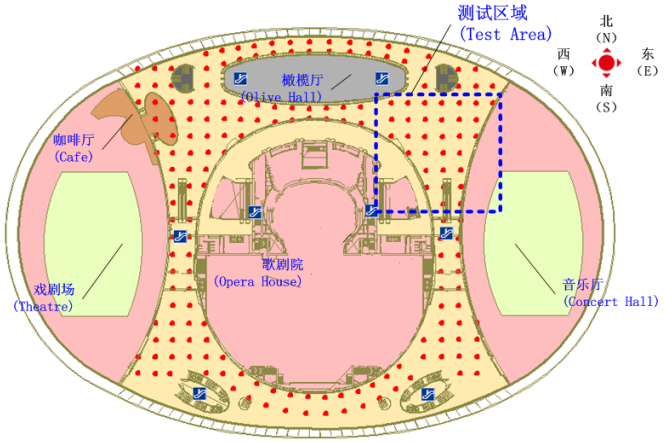

Figure 1. The red points are the reference points, The blue dash line represents test area.

B. δ selection

Data normalization first to define the range between 0 and 1. The size of the boundary region reflects the degree of roughness of the set in the approximation space, which was controlled by the values of parameter δ .

To optimize the parameter δ in NRS that control the size of the neighborhood, we have to do a pre-test. We test the range of each class which belongs to one feature. In our samples, see Fig.2, distances changes from 0 to 0.02. Using cumulative distribution function (CDF) to describe the range of class under a feature. 95% of the distances are within 0.015. We select this value as δ .

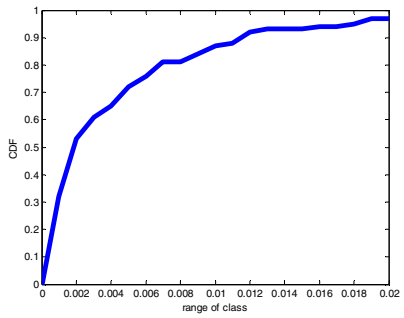

Figure 2. CDF of the range of class.

C. Experimental Results and Analysis

To access the superiority of our method, we conduct the experiments of NRS-Maxmean and NRS-PCA to compare with original Maxmean [9] and original PCA [13]. For fairness, KNN (K=1) is used to test the location performance.

Fig. 3 shows that effects of on the localization performance. The root mean square error (RMSE) decreases dramatically as the dimensionality increases at first and decreases slightly after then. The performance of NRS-PCA is better than NRS-Maxmean because NRS-PCA uses more information than NRS-Maxmean method. The number of selected APs is set to 8 in our experiment.

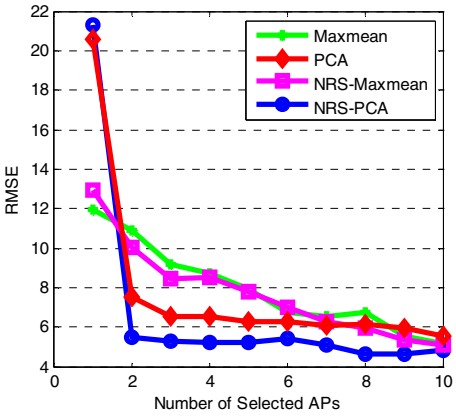

Figure 3. Effect of number of selected APs on the root mean square error (RMSE).

Fig.4 presents the cumulative accuracy within specified distances of different methods. Apparently, our methods obtain the better positioning accuracy than original methods. At error distance of 5 m, cumulative accuracy of Maxmean, PCA, NRS-Maxmean and NRS-PCA algorithm are 32%, 56%, 50% and 70% respectively. At error distance of 10 m, cumulative accuracy for these algorithms are 76%, 86%, 90% and 96% respectively. It means that NRS method is quite useful when applied to the original AP selection methods. Table 1 shows root mean square error (RMSE) of different approaches. Our methods (NRS-Maxmean and NRS-PCA) have dramatic performance improvement due to the pre-selection based on NRS. NRS-PCA works best among all methods, NRS-Maxmean has a comparable performance with the PCA method, while Maxmean performs the worst.

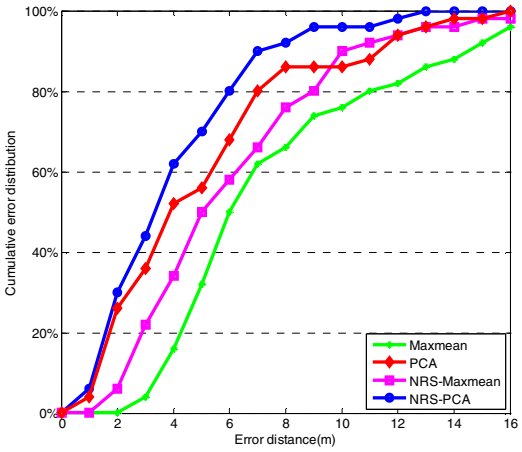

Figure 4. Location accuracy comparison.

TABLE I. PERFORMANCE COMPARISON

Algorithm	RMSE (m)	Location accuracy (%)	
		Within 5m	Within 10m
Maxmean	6.74	32	76
PCA	6.13	56	86
NRS-Maxmean	5.96	50	90
NRS-PCA	4.64	70	96

As can be seen, NRS-Maxmean and NRS-PCA perform better than the original methods. Thus, the proposed method based on NRS has the potential to convey more discriminative information and reduce the computational effort.

IV. CONCLUSION

In this paper, we propose a new AP selection method for indoor localization problem. The proposed method-based on neighborhood rough sets-offers a more discriminative way to select a small subset of available APs from lots of virtual APs in large scale buildings. Moreover, an adequate neighborhood δ has been chosen to control the boundary region. Experimental results also show that NRS can be used to further improve the performance of the original AP selection methods. The NRS-Maxmean and NRS-PCA perform much better than the original Maxmean and PCA methods.

ACKNOWLEDGMENT

The authors would like to thank the financial support provided by National High Technology Research and Development Program of China (863 Program) (No. 2009AA12Z324).

REFERENCES

[1] Akyildiz I F , Su W, Sankarasubramaniam Y, et al. A survey on sensor networks. IEEE Communication Magazine, 2002, 40(8) : 102–114

[2] Patwari N, Ash J N and Kyperountas S. Locating the nodes: Cooperative localization in wireless sensor networks. Signal Processing Magazine, 2005, 22(4): 54–69

[3] Sun G, Chen J, Guo W, et al. Signal processing techniques in network-aided positioning: A survey of state-of-the-art positioning designs. Signal Processing Magazine, 2005, 22(4):12–23

[4] Kushki A., Plataniotis N and Venetsanopoulos A N. Kernel based positioning in wireless local area networks. IEEE Transactions on Mobile Computing, 2007, 6(6): 689-705

[5] Brunato M and Battiti R. Statistical learning theory for location fingerprinting in wireless lans. Computer Networks, 2005, 47(6): 825–845

[6] Kaemarungsi K and Krishnamurthy P. Modeling of indoor positioning systems based on location fingerprinting. Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM'2004) : Vol 2, March 7-11, 2004: 1012–1022

[7] Sayed A H, Tarighat A and Khajehnouri N. Network-based wireless location: Challenges faced in developing techniques for accurate wireless location information. Signal Processing Magazine, 2005, 22(4): 24–40

[8] Li B H, Salter J, Dempster A G, et al. Indoor positioning techniques based on wireless lan. Auswireless Conference, 2006, CD-ROM pro: 13–16

[9] Youssef M, Agrawala A and Shankar A. Wlan location determination via clustering and probability distributions. Proceedings of the First IEEE International Conference on Pervasive Computing and Communications (PerCom ' 2003): March 23-26, 2003: 143–150

[10] Xu Y, Deng Z and Meng W. An indoor positioning algorithm with kernel direct discriminant analysis. Global Telecommunications Conference (GLOBECOM' 2010): Dec. 6-10, 2010:1-5

[11] Feng C, Au W, Valaee S, et al. Received signal strength based indoor positioning using compressive sensing. IEEE Transactions on Mobile Computing, 2011, PP(99): 1–12

[12] Yiqiang, C., et al., Power-efficient access-point selection for indoor location estimation. Knowledge and Data Engineering, IEEE Transactions on, 2006. 18(7): 877-888.

[13] Shih-Hau, F. and L. Tsungnan, Principal Component Localization in Indoor WLAN Environments. Mobile Computing, IEEE Transactions on, 2012. 11(1): 100-110.

[14] Pawlak Z.Rough Sets, Theoretical Aspects of Reasoning about Data.Dordrecht:Kluwer Academic,1991.

[15] R. Jensen and Q. Shen, "Fuzzy-rough data reduction with ant colony optimization," Fuzzy Sets and Systems, 2005, 149(1): 5–20.

[16] Hu, Q., D. Yu, and Z. Xie, Neighborhood classifiers. Expert Systems with Applications: An International Journal, 2008. 34(2).