Received: March 3, 2021. Revised: March 25, 2021.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

An Ameliorated Ensemble Approach for IoT Resource Feature Selection Based on Discriminating and Service Relevance Criteria

Vandana Charaliparambu Pathayapuram[1]* **Ajeet Annarao Chikkamannur[2]**

[1]*Visvesvaraya Technological University, New Horizon College of Engineering, India*
[2] *Department of Computer Science and Engineering, R.L. Jalappa Institute of Technology, India*
* Corresponding author's Email: vandana.hareesh@gmail.com

Abstract: Internet of Things (IoT) paradigm merges the physical world into digital world with proliferation of smart solutions in various domains. These smart solutions leverage the sensing and actuating power of heterogeneous IoT devices. The user applications must discover the right IoT resource at the right time. To serve this objective, the most important attributes of these diverse IoT devices must be identified and managed by network management systems (NMS) for effective discovery. In this paper, an ameliorated feature selection approach is proposed which selects minimum number of discriminating and service relevant features of multiple diverse IoT devices. Proposed feature selection technique has two phases where the features are ranked based on discriminating capability. Minimal subset of discriminating features is then chosen and ensemble with service relevance features. The proposed model is validated with dataset constructed from annotated IoT resources based on our previous work. Analyzes is carried out by comparing with existing state of art hybrid feature selection techniques and the proposed method has outperformed in terms of accuracy, quality and size of feature set selected. Further, these selected feature set derived from the proposed feature selection method is used to enrich IoT protocol, Constrained Application Protocol (COAP) CoRE Link format. This semantic enriched COAP protocol is used for IoT resource discovery on COOJA simulator. The results show that enriched COAP based discovery diminishes the discovery time by 33% compared to traditional COAP. Experimental work is carried out to showcase the effectiveness of the proposed feature selection model in terms of resource discovery success rate, query response time, service type detection accuracy.

Keywords: Feature selection, Semantic, IoT, Similarity matrix, Service relevance, COAP, Discriminate, Semantic.

1. Introduction

The Internet of Things (IoT) [1] is attributed as the revolutionizing paradigm connecting our physical environment to digital space. With the plethora of heterogeneous IoT devices, various domains like smart homes & cities, Intelligent Transport system, e-agriculture, Industrial IoT (IIoT), healthcare and other enterprises are leveraging the power of IoT to provide reliable and timely user applications and management solutions. By 2025 [2], it is estimated that that around 75.44 billion devices comprising of sensors, actuators, computing units like cameras, power switches, door locks, microphone, speakers, temperature & gas sensors, motion sensors, household appliances would be engendering our society and lives.

These smart devices comprising of physical components and facilitating services are termed as 'IoT resources'. With the proliferation of multiple heterogeneous IoT resources, it is an inevitable requirement for the operators and network management system (NMS) providers to connect and manage the right resource and provide the uninterrupted services in the user space. Discovering [3] the most suitable resource based on its capabilities at the right time and right context is a need of time.

IoT protocol stack envisions multiple application layer protocols like COAP [1], MQTT [1] to discover and provide IoT resources. However, for efficient resource discovery of these multiple, resource constrained, heterogeneous IoT resources, semantic

based enrichment of application protocols has been proved to be efficient. Our previous work S-COAP [4], presents the semantic enrichment of CORE Link format and studies its efficient discovery of IoT resources. Semantic description of IoT resource [5] and IoT resource component model was proposed and aligned to various ontology like SSN [6], IoT-A [6]. Semantic label mapping to the text words extracted from the IoT device vendor specifications facilities a uniform representation of IoT resource and helps to identify the resource capabilities and uniform access.

The semantic annotation [7] of IoT resource augmented by ontology is performed based on product specification. IoT resource annotation presents several attributes representing its physical device components, deployment units, services offered, quality of services and other describing features. In order to semantically enrich protocols for resource discovery features rich in identifying the resource and its capabilities and discriminating multiple heterogeneous devices needs to be selected from the numerous attribute sets. NMS providers can manage multiple IoT resources based on these features and can operate various network management services. Feature selection [8] selects an optimal subset of features from the feature space based on the selection criteria reducing the curse of dimensionality and improving the time and space complexity of the algorithms. Finding the global optimal feature set preserving the independency and at the same time combined effect is an NP hard problem and growing areas of research.

In this paper, we design a methodology to select minimum number of attributes from IoT resource annotated feature space [7], in order to assist in efficient registration and discovery process of these resources. Since IoT devices are resource constrained, minimum and optimal features need to be chosen which are both discriminating at the same time identifying its capabilities. Contributions made through this paper include:

- To design a structured ensemble attribute selection framework from IoT resource annotated feature space based on criterion of maximum discrimination, minimum redundancy, and maximum capabilities relevance discovery.
- Construct the consistent, optimized pair wise-comparison matrices of attribute discriminating ability and compute the feature score.
- Methodology to select the minimal set of discriminating attributes, removing the redundant features. Identifying service relevant attributes based on mutual information criteria restricted by Minimum description Length.
- Ensemble feature set used to enrich COAP protocol and verify the discovery effectiveness with COOJA[24] simulator study.

The related work in this area is described in Section2. Section3 represents the proposed ensemble framework; methodology for discriminating ability based feature selection is presented in Section 4. Section 5 describes the service relevant attribute selection. Section 6 discusses the performance evaluation of the overall structured ensemble feature selection framework Section 7 briefs the conclusion and future work for resource discovery based on proposed feature selection ensemble.

2. Literature survey

With the proliferation of heterogeneous IoT devices in market, network management solutions find it challenging to choose most significant attributes describing these resources. Semantic enrichment [4] of IoT application protocols like COAP can enable efficient resource discovery of the IoT devices. However, finding the minimal set of attributes for heterogeneous devices to discriminate and service relevance is a non-trivial task. Feature selection techniques helps in selecting these important relevant features with respect to the management operation. In Machine Learning, feature selection[8] can be performed either by filter technique, wrapper or embedded approaches.

Under filter approach, the features can be ranked based on various statistical measures or information theory based concepts. Most popular scoring techniques include Laplacian score (LS) [9], feature correlation[8], Pearson correlation coefficient (PCC) [9] and Fischer score[9]. The scored features can be ranked and then top scoring features can be used to improve accuracy of the classification or clustering models. However in case of filter approach, the relationship between features may not be considered for feature selection. Hence redundant attributes may be selected.

According to Fisher Criterion [9], attribute variance would be small for the data points associated with same class, while it will be large for data points associated with different classes. However, with low sample size, performance of Fischer Criterion may degrade due to computation of mean and variance with biased values.

In [10], Relief based feature approach is reviewed which is based on proxy statistics metric. Relief variant, ReliefF the comparison with nearest neighbours is conducted, where the number of nearest neighbours is a user input.

Using wrapper approaches, feature subsets can be chosen which satisfy the criterion leading to an optimal feature set generation. Popular approaches include sequential forward selection (SFS) [11], sequential backward selection (SBS) [11] where the feature is add sequentially to empty set or eliminated from full feature set based on the accuracy outcome of the classification model used in the approach. Compared to filter approach, it considers the relation between class label and features but results can be biased based on the classifier used in the model.

To leverage the advantage of both filter and wrapper approach hybrid approaches are proposed. Mutual Information (MI) [12] based information theory concept is used for ranking the features. MI is computed between a feature and all its neighbour features. Here algorithm selects the feature with increased MI with class label and decreased MI with other features. It is a greedy approach with increased computational cost. In the hybrid approach [13], MI based filter technique and Recursive Feature Elimination (RFE) based wrapper technique is adopted. However, RFE requires the user to input the limiting count for the feature set size. Genetic Algorithm based feature selection is performed in [14, 17] where the optimization in feature selection is achieved. In [15], validity index metric is used as a parameter for ranking the features followed by wrapper technique. However, the validity index metric is local bound and does not consider global variability behaviour. Feature selection is performed by pair wise evaluation on a given dataset in [16] and uses it as a pre-qualifier in terms of quality and relation between the features. Feature subset selection algorithms are modified based on this pre-evaluation result. In the work [18] distance based metric is used to rank the features followed by feature subset wrapper construction based on bootstrapping technique to find the optimal subset. Unsupervised feature selection[19] is proposed by clustering the features based on similarity index. However, the input for cluster count needs to be provided by the user in case of KMeans algorithm.

Most of the recent research work[20,21] includes identifying the IoT device deployed using the traffic analysis, in security domain. Feature ranking for IoT device classification is performed on the device traffic based on statistical measures [20]. The feature capable in predicting the class label of the classification model is chosen. However, with low size of sample set, training the classification model may degrade the performance. The feature selection from IoT device traffic is performed based on misclassification loss [22] as risk cost.

The current state-of art of IoT resource identification depends heavily on the supervised data for training the model. Features are extracted from the network traffic, affected by the quality of the collected data(encrypted traffic). In our proposed model, feature selection is carried out on the IoT annotated data extracted from vendor specification [7]. To assist IoT resource discovery, proposed model targets feature selection based on discriminating capability of device, service relevancy. Also minimal subset is derived so that resource constraints IoT protocols can benefit from these identified features and enhance discovery process.

3. Literature survey

The attributes of an IoT resource is annotated based on the IoT resource ontology discussed in [5,7]. Attributes are broadly divided [5] into Device Entity, Service entity, Telemetric entity and Service Composition entity. Each entity component is a collection of all the describing attributes. Since the IoT devices are resource constrained, their corresponding protocols are also resource constrained. Usage of all the attributes of a device for the discovery process leads to exhaustive memory, computation cost which is not feasible.

Thus attributes which exhibit maximum discriminating ability between the resources, and depict service relevance criteria must be selected for efficient resource discovery process as shown in Fig. 1. The discriminating ability of the attribute would help to uniquely identify the IoT Resource. The attributes DeviceID, ServiceType, Brand, Model, Mounting type, etc possess the discriminating ability. The service relevance attributes include the Quality Of Service (QoS) attributes namely, Precision, Accuracy, Resolution rate, Pixel resolution. These service relevance attributes are specific to the servicetype attribute.QoS attributes when extracted from vendor specification possess definite values which wont exhibit discriminating ability amongst the devices offering same servicetype. After deployment of the IoT resources in runtime environment, these attributes must be monitored by the NMS system. These attributes possess dynamic values and are important to define the capabilities based ranking of the devices. Hence discovery process need the service relevant attributes for automatic discovery of the most pertinent devices and

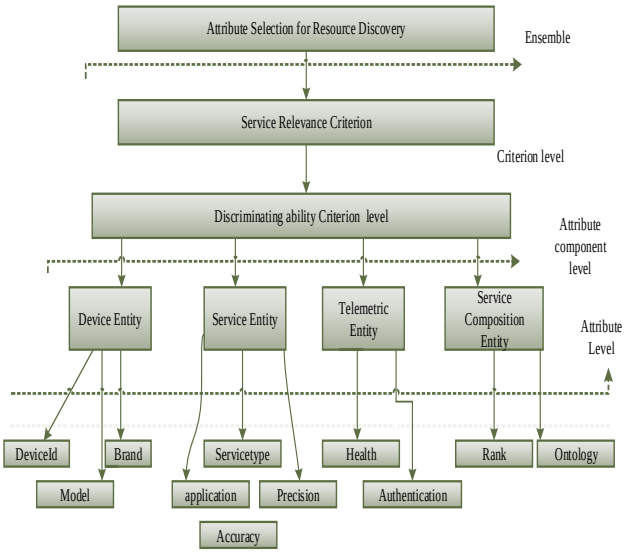

Figure. 1 Feature selection for IoT resource

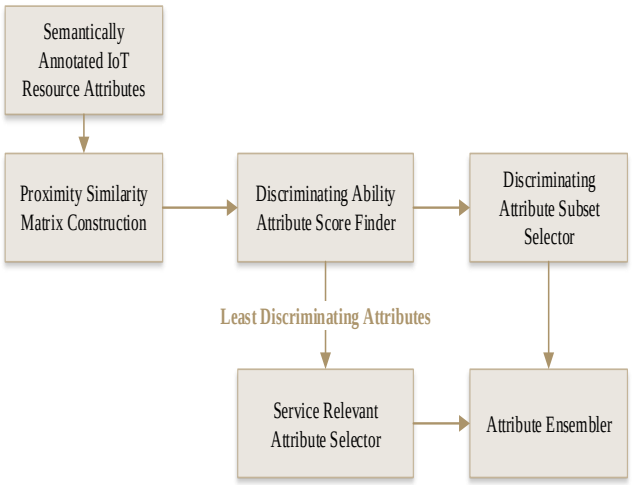

Figure. 2 Attribute selection based on discriminating and service relevant criterion

for recovery process incase of faults in the network.

The ameliorated ensemble approach for feature selection is shown in Fig.2. The attribute dataset[7] is processed to construct the Feature Proximity similarity matrix (SM) based on pair wise comparison of the attribute values.SM is employed to compute the discriminating score for the attributes.

The score finder returns the ranked feature set capable of discriminating the IoT resources. The most discriminating feature, device id can be identified using this phase. The ranked feature set is then employed to find the minimal optimal subset of discriminating attributes, removing the redundant attributes. The remaining attributes (except the discriminating subset) is subjected to service relevance finder process. Service relevance based ranked attributes are augmented to the final attribute set, limited by Minimum description length (MDL). MDL considers the IoT protocol resource constraint. This amalgamation results in minimal set of attributes that possess maximal discriminating, minimal redundancy and service relevant features.

4. Discriminate criteria based IoT resource feature scorer

The objective of this procedure is to obtain a score for each attribute, which is indicative of its discriminatory power to differentiate between IoT resources and then order attributes based on the score. For a given feature, if the characteristics of two IoT resources are not equal (Proximity Measure), then we can likely discriminate between the two IoT resource using that attribute. For the pair of IoT resources that are similar, the values of attributes should be close. Feature selection technique comprises of finding attribute dimensions along which the value pairs are very close for same IoT resource, however very far for different IoT resources.

4.1 Proximity measure

We define a function P(X), the proximity measure x,yε X such that

$P(x,y) \geq 0$
$P(x,y)=P(y,x)$
$P(x,y) =P(x,x)$ iff x==y

This proximity measure is based on pair wise attribute similarity. $P_{ij}= [d1, d2\ldots dn]^T$, di can be value between 0 or 1 depending upon if the degree of similarity or dissimilarity between pair of attributes values. The metrics used for the computation of the proximity measure includes Manhattan distance[9] for numerical data, Jaccard index[9] for categorical data. In case of nominal values, they are converted to continous values and then proximity measure is applied.

Manhattan Distance is given in Eq. (1) as where x_i and y_i represent the features of device i.

$$P(x,y) = \sum_{i=1}^{n} |x_i - y_i| \quad (1)$$

Jaccard Index is computed for categorical value list for features X and Y of a particular device in Eq. (2)

$$P(x,y) = \frac{|X \cap Y|}{|X \cup Y|} = \frac{|X \cap Y|}{/X/+/Y/-|X \cap Y|} \quad (2)$$

Distance between nominal attributes is computed as Eq. (3), where p is the total number of attributes and m is the number of matching /similar attributes

$$P(x,y)=\frac{p-m}{p} \tag{3}$$

For one attribute and n IoT resources, *n * (n-1) /2* pair wise comparison is the requirement. Proximity Similarity Matrix (SM) represents the distance similarity between attribute values of pair IoT resources. Proximity measure $PM_{x,y}$ is computed as Eq (4).

$$\begin{aligned} PM_{x,y} &= 0 \quad , \quad P(x,y) == 0 \\ &= P(x,y), Otherwise \end{aligned} \tag{4}$$

The Boolean element is pair wise comparison on attribute f, between IoT resource x and y as in Eq (5).

$$\begin{aligned} SM_{fx,y} &= 1, \quad PM_{x,y} > \theta\ (threshold) \\ &\quad 0 \quad\quad PM_{x,y} = 0 \end{aligned} \tag{5}$$

The number of the pair wise comparisons would be n*(n-1)/2 would be optimized by applying the transitivity rule.

4.2 Proximity similarity matrix with transitivity rule

As per the transitivity rule, for a scalar value 'a', indexes i,j,k ,if $a_i = a_j$ and $a_j = a_k$ then it implies $a_i = a_j$

Applying the transitivity rule, for a particular feature f, consider any three IoT resources x,y,z , Proximity Similarity Matrix (SM) is given as

$$If\ SM_{x,y} = SM_{x,z}\ then\ SM_{y,z} = SM_{x,z} \tag{6}$$

Hence according to Eq. (6), SMy,z need not be computed as it will be same as SMx,y or SMy,z. If SMx,y is not same as SMx,z , then SMy,z needs to be computed. Applying this rule, in best case where transitivity rule applies, only n-1 comparisons would be needed. Also, in the case of missing values, this rule can be applied to fill the missing values.

Feature Proximity Similarity Matrix (PSM) would comprise of the pair wise Proximity Similarity Matrix (SM) as shown in Eq. (7).

$$PSM = \begin{bmatrix} NA & NA & NA & & NA & NA \\ SM_{f_{2,1}} & NA & NA & & NA & NA \\ SM_{f_{3,1}} & SM_{f_{3,2}} & NA & & NA & NA \\ . & . & . & & . & . \\ . & . & . & \dots & . & . \\ . & . & . & & . & . \\ SM_{f_{i,1}} & SM_{f_{i,2}} & SM_{f_{i,3}} & & NA & NA \\ SM_{f_{j,1}} & SM_{f_{j,2}} & SM_{f_{j,3}} & & SM_{f_{j,i}} & NA \end{bmatrix} \tag{7}$$

The summation of the PSM scores fsmf in Eq. (8) would provide the discriminating index for each attribute from the range (1..k) where k is the number of attributes for a group of similar IoT resources.

$$fsm_f = \frac{\sum_{i,j;j>i}^{n} PSM_{f_{i,j}}}{n(n-1)/2} \tag{8}$$

For K features of an IoT resource, the range of symmetry scores achieved would be in less than 1 as shown in Eq. (9)

$$\begin{aligned} &fsm_{f1,} fsm_{f2,} fsm_{f3,} \dots\dots\dots fsm_{f,k} \\ &where\ 0 \le fsm_{f,k} \le 1 \end{aligned} \tag{9}$$

Based on the score, top high valued attributes which have high discriminating power are considered. Algorithm 1 FSM, performs the discriminating score computation based on Feature Similarity Matrix.

Algorithm 1 FSM: Discriminate_Score_Finder

INPUT: N devices with F Attributes, N∈{ 1,2,3,...n} F∈{1,2,3,..f}

OUTPUT: Subset of F Attributes $f_{discriminate}$

1. for k ∈{1,2,..f} and k>=f do
2. Sim=[ ]
3. Compute the proximity measure $PM_{x,y}$ using (4)
4. Choose threshold θ using Algorithm 2
5. Map $PM_{x,y}$ to Pairwise Boolean Similarity $SM_{fx,y}$ using (5).
6. Apply (6) to Construct Feature similarity matrix FSMs in (7).
7. Initialize feature score $\psi_f = 0$
8. for i,j ∈{1,2,..n} and j>i do
9. $\psi_f = \psi_f + PSM_{f,,i,j}$
10. end for
11. Sim←Sim + { ψ_f }
12. end for
13. S←sort(Sim)
14. D←sort(Sim, reverse_order)

15. $\forall f \in \{S \cup D\}$
16. $Fscore_f = (|0.75S_f - D_f|) / (|0.75S_f + D_f)$
17. Rank the features according to Score
18. return the feature with highest score as the most discriminating feature and the remaining leading m features

Attributes with high similarity values need to be considered for its closeness affinity within same devices. So the attributes which have high probability in both sets are highly preferable with high discriminating power. The attribute with highest discriminating ability would be the DeviceID.

Algorithm 2: Threshold Setter

INPUT: Proximity similarity matrix SM_s of N devices

OUTPUT: Threshold θ

1. Find the minimum value (>0) of each row(i)of SM_s and store as min_val_i , where $1 \leq i \leq N-1$
2. Set the first percentile of these N-1 values of min_val_i as θ
3. Return θ

The Algorithm 2 provides the threshold value for constructing PSM based on SM value. The first percentile (.25) of the minimum values is taken to compute the threshold value. Based on this threshold, PM values are converted to Boolean value in PSM.

4.3 Minimal discriminating feature subset selector

The IoT resource attributes are ranked according to their discriminating scores and given input to proposed Minimal Discriminating Feature Subset Selector (DFSS) algorithm.

In case of exhaustive search for feature subset selection, without considering any feature ranking, finding an optimal minimal set is a NP hard problem.

In Sequential Forward Selection (SFS)[11] , the attributes are selected sequentially one by one and added to the initial empty feature set. If the selected attribute improves the discovery criteria, the attribute is selected to be part of final set. This is repeated till the unique IoT resource is discovered by the test query.

In case of Backward Feature Elimination (BFE)[11], the entire attribute set is taken and one by one, elements are checked against the discovery criteria. If the presence of the attribute does not improve the criteria or degrades it, the attribute is eliminated from the feature set. This process is repeated till unique device id is returned.

In both SFS and BFE, the redundant features may be a part of the subset selected. Hence to overcome the limitation of exhaustive search, the ranked features are considered in order. Also backtracking of features is performed to overcome SFS and SBE limitations.

DFSS algorithm takes the attributes from highest scored to least in the order. The highest scored feature, DeviceId must be uniquely retrieved based on the minimal set of discriminating features. Algorithm 3, at each step sequentially picks the next feature in the order and checks the discovery result. If the response improves, (number of retrieved similar devices reduces), the feature would be included in the feature set. This process is repeated until the discovery process (Algorithm 4) returns only one unique device. In order to reduce the redundant features which can occur during the combination, backtracking is performed to verify if the discovery response is same after removing the redundant features.

Algorithm 3 DFSS: Discriminating Ability based Minimal Feature Set Selector

INPUT:Feature Set ordered based on discriminating ability score $D=\{f_0, f_1, f_2,..., f_n\}$. f_0 is the most discriminating attribute, device id. Remaining Set $D=\{f_1,f_2,f_3,...f_n\}$ has n-1 features including service_type. MAX= Dataset Size.

OUTPUT: minimal set of discriminating features

```
1.  Initialize selectlist={} prevdiscover=MAX
    finallist= {}
2.  for i=0 to n-1 do
3.      selectlist.append(D[i])
4.      currdiscover = Discover_Device (selectlist)
5.     if currdiscover == 1 then
            finalist=selectlist
7.      else if currdiscover < prevdiscover
            if i! = n-1 then
                prevdiscover= currdiscover
            else
                finallist= selectlist
8.       else
            remove D[i] from selectlist
9.      for j=0 to length(finalist)-1
            minimal= Discover_Device (finalist-dj)
          if (minimal==Discover_Device (finallist))
                finallist=finallist-dj
   10.    return finallist
```

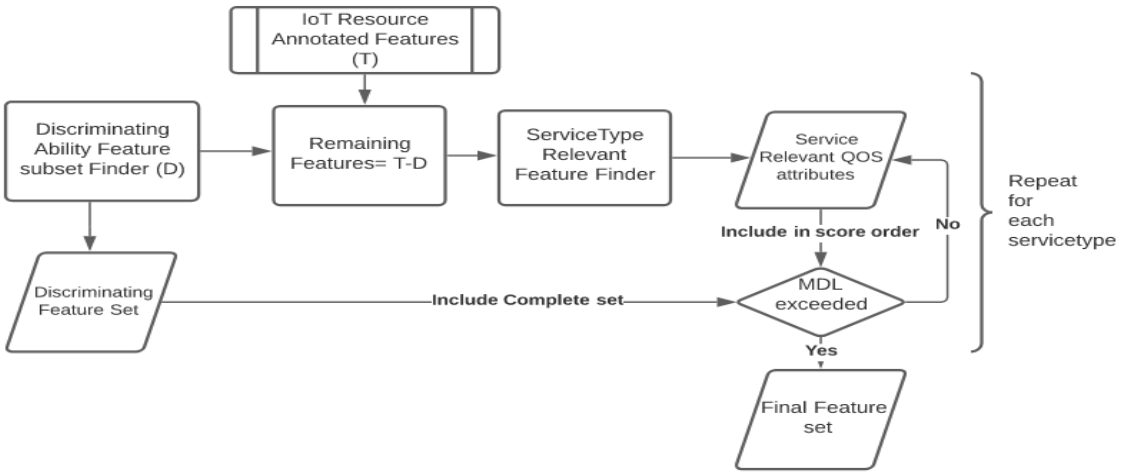

Figure. 3 Service type relevant attribute selection

Algorithm 4 performs the deciding criteria for minimal feature selection. The SPARQL[22] query is executed for different values from the feature set such that it covers maximum number of tuples.

Algorithm 4: Discover_Device

INPUT: List of discriminating features D, TestQuery S with attribute $\{s_1,s_2..s_p\}$
OUTPUT: Number of devices discovered

1. For each $q_i \in D$ do
   $q_1[value]==s_2[value] \wedge q_2[value]==s_3[value] \wedge q_3[value]==s_4[value] \wedge \ldots.\wedge q_3[value]==s_4[value] \wedge q_p[value]==s_p[value] => RS \{ rs_1, rs_2, rs_3,.. rs_k \}$
2. Return RS

The SPARQL query incrementally tests the discriminating behaviour on the RDF graph[22] of the device data set as given below:

```
qres = RDFGraph.query("""SELECT ?s
WHERE {
        %s
        ?x
<info:discovery/iot_resource/device_id> ?s }""
" %feature)
```

5. Service type relevancy based IoT resource feature selection

After the first phase, the minimal set of features capable of discriminating the IoT resources is selected. Mutual Information (MI) [11] is used to find the relevance of the feature with respect to the servicetype.

MI is measures the amount of information one random parameter provides about another parameter. It takes a value of zero when there is no relationship between parameters and a positive value for strong relationship existence. MI measures the entropy of the system, ie the measure of disorder in the system. MI computes the relevance of the attribute with respect to servicetype (class label). Calculating the MI for an attribute is performed by computing the entropy of the servicetype for the entire dataset and reducing the conditional probabilities for each possible value. The features would be scored based on service relevancy. The QoS attributes which are unique to the device service type would be selected. As depicted in Fig 3, the minimal feature set qualifying the discriminating criteria is removed from the initial feature set. This is further given as input for service type relevant feature selection process. In Information theory, the Mutual Information (MI) between two random parameters is given by Eq. (10)

$$MI(A:B) = \sum\sum P(a,b) log \frac{p(a,b)}{P(a)(b)} \quad (10)$$

Where p(a,b) is the joint probability distribution function of the attributes A and B. p(a) and p(b) are the marginal probability distribution functions for A and B. Properties of MI is given below

$MI(A;B)=MI(B;A)$

$MI(A;B)=H(A)+H(B)=H(A;Y)$

Where H(A) is the entropy of the random parameter A and represents the uncertainty about this parameter, H (A; B) is the joint entropy of the parameters A and B given by Eq. (11) and Eq. (12)

$$H(A) = -\sum_{a \in A} P(a) \log P(a) \quad (11)$$

$$H(A;B) = \sum_{a \in A} \sum_{b \in B} p(a,b) log p(a,b) \quad (12)$$

Algorithm5 SRFS: Service Relevance Feature Selector

INPUT: Remaining Attributes with least discriminating ability S={f_0,f_1,..f_m} C=Service_type

OUTPUT: Service Relevant feature set

1. for i =0 to m-1
2. compute MI(fi,C) using (10)
3. Service_rel_list=Reverse_Sort(MIC(fi,c))
4. Return Service_rel_list

The target class label is considered as the ServiceType attribute. So, the relevant features towards the Servicetype are computed using Mutual Information (MI) as depicted in Algorithm 5. The relevant features in decreasing order of their relevance are considered. These features would represent the quality of service attributes.

5.1 Ensemble feature set

The final ensemble of the features would be the minimal feature set of discriminating scores and along with service relevant feature set. The number of features selected would be limited by Minimum description length (MDL).Minimum Description Length (MDL) is defined by the total number of features encoded in the IoT protocol header.

6. Result discussion

6.1 Experimental study

The annotated IoT resource feature set from the vendor specifications based on our previous work [7] is taken as the input feature set (data set), https://www.kaggle.com/vandanacp/iotresourceannotationfromspecfication

The proposed algorithms PSM, DFSS, SRFS are implemented in Python 3. Berkeley DB (BDB) with Rdflib5.0 library is used for feature store and the rdf format conversion to query. The source code is checked in and available at https://github.com/vandanacp/IoTResourceDiscoveryProject

The device annotations in rdf is uploaded as dataset for research community. https://www.kaggle.com/vandanacp/iotresourceannotationfromspecfication?select=IoTResource.rdf

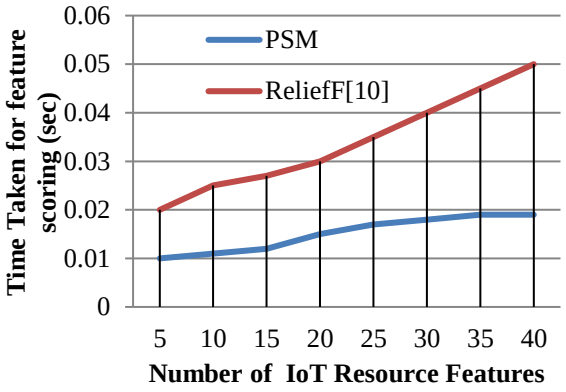

Figure. 4 Performance comparison of discriminating feature score finder PSM & ReliefF[10]

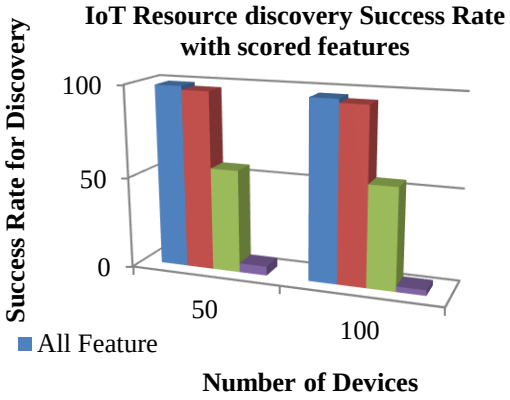

Figure. 5 Success rate with high scored feature

The proposed discriminating feature scorer technique is compared with ReliefF[10] in terms of computation time for computing the scores. The Fig. 4 depicts that proposed PSM outperforms reliefF by 20% , as the number of features increase.

This is due to the fact that ReliefF first converts all features to continuous data type which is time consuming.

The features are ranked based on PSM discriminating scores. As shown in Fig5, top 75% of these ranked features give same performance as the complete unordered feature set in terms of DeviceID based resource discovery. Results infer that the bottom 25% features are non-discriminating and possess same values for similar service providing IoT devices. Hence they are the service relevant features like QoS attributes which would be specific to a particular servicetype. So, proposed PSM is capable of inferring most discriminating attributes and service type attributes (QoS) attributes in ranked order. As shown in Fig 6, the proposed DFSS algorithm is compared with Backward Feature elimination (BFE)[11] and its counterpart SFS[11] for minimum feature set selection time. The query

Table 1. Discriminating feature set selection

Algorithm	Number of Initial Features for a specific Service Type	Number of discriminating features selected
SFS[11]	40	6
BFE[11]	40	6
Proposed DFSS	40	4

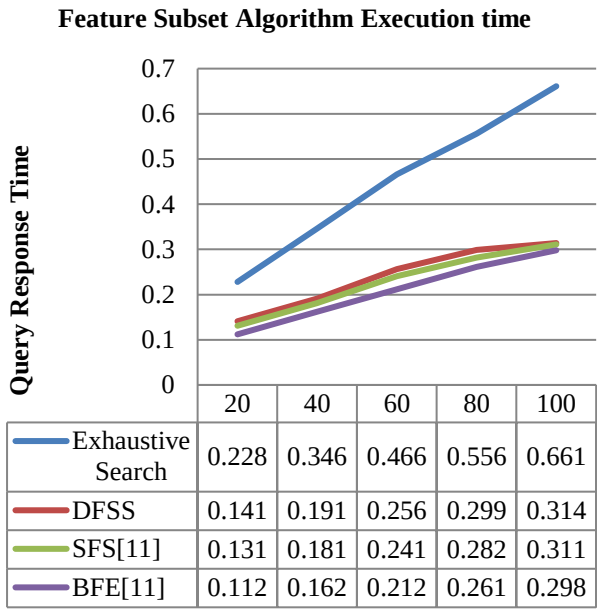

Figure. 6 Query response time comparison for feature subset algorithms

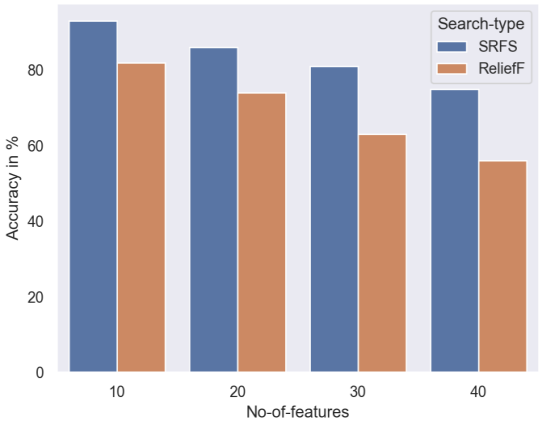

Figure. 7 Service type accuracy comparison

time for proposed DFSS is proportional to the number of features increasing from 20 to 100. The query time for DFSS is comparable with SFS and BFE, including time incurred for redundant feature set removal using backtracking. Exhaustive Search takes twice the amount of time compared with SFS, BFE and DFSS.

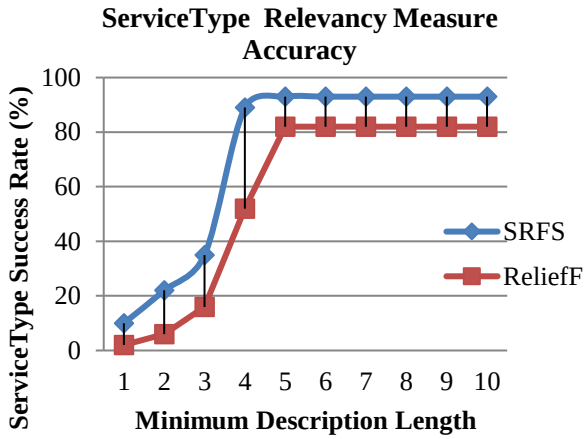

Figure. 8 Success rate vs. minimum description length

The number of features selected by DFSS has successfully eliminated 12% redundant features when compared with SFS and BFE as shown in Table 1. ServiceType relevant attribute identification using proposed SRFS technique is compared with ReliefF[10] in terms of accuracy. Confusion matrix computation, followed by accuracy computation is performed. Fig.7 depicts that SRFS gives accuracy improvement of 20% compared with ReliefF. ReliefF considers the nearest neighbours for similarity. ReliefF was executed using input parameter 10, as found optimal during our study. For a servicetype "Get_image", the final feature set would comprise of Device_ID,ServiceType, Device_Name,Mounting_Type,Model,weight,Video Resolution,VideoModes.

In fig8, proposed SRFS and ReliefF[10] success rate is compared with increasing Minimum Description length. Success Rate is defined as Eq. (13)

$$Success\ Rate = \frac{Number\ of\ Correct\ ServiceType\ Retrievals}{Total\ Number\ of\ Queries} \quad (13)$$

It is observed from the fig8 that beyond MDL length of 4, SRFS success rate is constant, hence MDL of 4 is ideal with the considered dataset. However, ReliefF success rate suggest an MDL of 5. Since these features are further used for IoT resource constrained protocols, MDL must be as small as possible to reduce network bandwidth usage. Proposed SRFS is capable of inferring the least MDL as possible to be 4.

We compared the various phases of our proposed feature selection model and found to outperform in terms success rate, accuracy, query response time.

Table 2. Comparison of proposed model with previous studies

Compared Models	Techniques	Features selected
MIFS-ND[12]	MI between features and MI feature and class label	14
Hybrid[13]	MI + RFE	12
LMFS[18]	Distance metric+ bootstrapping	12
Proposed Model	FSM + DFSS+ SRFS	8

Table 3. Classification measure with respect to various feature selection methods

Method	F1-Score	Precision	Recall	Accuracy
MIFS-ND[12]	93.13	93.71	93.13	93.93
Venkatesh [13]	92.13	93.11	92.13	92.93
LMFS[18]	91.13	91.12	91.13	92.01
Proposed Model	**96.13**	**96.71**	**95.13**	**96.31**

6.2 Other compared work

In this section the proposed ensemble approach is compared with current state-of-art hybrid approaches. The dataset [7] is used for comparison study which has IoT devices with 40 features. The proposed model selects minimum number of features from the given dataset of annotated features of IoT resources. Sub set feature size is highest in [12] as it computes the MI between the features (pairwise) and MI between features and serviceType(class label). It incrementally adds feature with high label MI compared to pairwise MI. Our proposed model outperforms by average 15% in feature subset size criteria as shown in Table 2.

The quality of feature subsets derived by various compared models is evaluated with KNN[8] classifier. The KNN classifier learns based on distance metric with its k neighbours. Test data is classified to multi classes representing the IoT resource Service_Type attribute. The value of k is set to be 5 in python sklearn package based

implementation. The confusion matrix based classification report represents True positive (TP), True Negative (TN), False positive (FP) and False Negative(FN) values. Table3 shows that our proposed method achieves more accuracy (96.31%) and Precision (96.71%) compared to other methods in comparison.

The selected feature set possess service relevance attributes to assist in servicetype classification carried out with KNN classifier with improved results compared to existing models.

6.3 Simulation based performance study of selected features by enriching COAP protocol for discovery

In the previous section, the selected features are evaluated with classifier with respect to Service_type class label. However, selected features needs to be evaluated for its effectiveness in IoT device discovery process. Hence we evaluated the fitness of these identified features by using COAP protocol in COOJA[25] simulator environment. COAP is request/reply constrained application protocol and maintains a resource directory (RD) for all registered IoT devices. As discussed in our previous work [4], we modify(enrich) COAP CoRE Link[24] format with these selected features. COAP RD is modified to retrieve devices using identified features.

6.3.1. Design of methodology to retrieve objects in COAP resource directory (RD)

Discovery of most pertinent resource based on static and dynamic attributes.

- Resource Directory RD maintains Static (S) and Dynamic (D) Attributes of IoT resources $\{R_1,R_2..R_n\}$ based on their values $S=\{s_1,s_2,s_3,..,s_n\}$ $D=\{d_1,d_2,d_3,\ldots d_m\}$
- Given a User request (Q) comprising of requested IoT resource attributes $Q= q_1,q_2,..q_i$ where $qi \in \{S,D\}+$
- Initialize search criteria with initial Dynamic Attribute based on user request Q
- If matching value of Dynamic Attribute for IoT resource found in search space, then
    - Map Dynamic Attributes to corresponding Static Attribute, $\{d1d2..dn\}\rightarrow s1$
    - Return the discovered IoT resource R_i
- Else
    - Include the next Dynamic Attribute ‘di’ in the search criteria and perform the search.
- Repeat until corresponding Static Attribute is found, IoT resource is discovered.

Table 4. Simulation parameters

Parameter Name	Value
Radio medium	Unit Disk Graph Medium
Transmission range	50 m
Bit rate	250kbps
Sensor type	Sky-mote
MAC layer/ Radio duty cycling (RDC)/Adaptation Layer	ContikiMAC / CSMA/CD 6LoWPAN
Routing Protocol	IPv6/RPL
Number of nodes	43
Network	350m * 350m
Simulation Time	1500 seconds

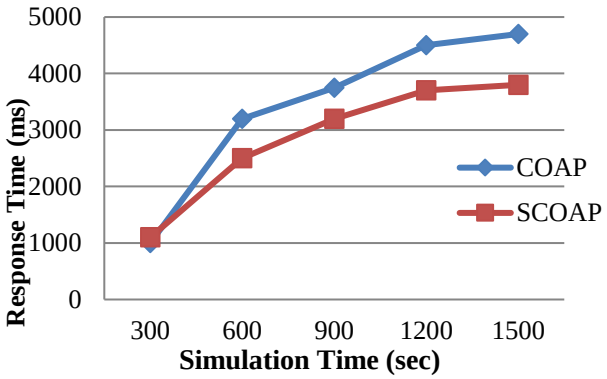

Figure. 9 Comparison of response time with COAP and modified COAP

A simulated testbed with 43 COAP supporting sky –motes were deployed. 20 Client instances were initiated. COAP registration process is performed with minimal set of features derived from our proposed model. Table 1 refers the various simulation parameters configured for COOJA simulator on Contiki OS.

Fig.10 depicts that query response time comparison between traditional COAP and semantically enriched COAP (SCOAP) with our derived feature set. SCOAP performs 33% better in discovery time compared to COAP. Initially, at time 300 seconds, resource registration phase commences. As simulation time progresses, discovery response time is recorded for the user queries run in batch of 5. Thus the poposed methodology selects the feature set which has good discriminating ability to discover deviceid uniquely as shown in simulation study.

7. Conclusion and future work

In this paper, we presented an ameliorated ensemble approach for IoT resource feature selection which can assist in efficient resource discovery and network management operations. In the proposed feature selection model, selection criteria are based on finding the most discriminating attributes and features relevant for service type identification and its quality. Feature set extracted from device vendor specifications are employed [7]. The effectiveness of proposed discriminating feature scorer is compared with other filter approach. Proposed DFSS algorithm derives minimal set size compared with other wrapper algorithms. This ensemble approach outperforms in terms of classification accuracy in comparison with existing hybrid feature Also, the selected features were used to enrich COAP protocol[4] and was used in COOJA simulator based study. Resource discovery is carried out in simulated environment with traditional COAP and enriched COAP with selected features to augment the result validation. Discovery time improvement shows the success of the selected feature set derived from our proposed model.

In future, we would propose a framework for ranking the IoT resources. Semantically enriched COAP protocol with the selected feature set would be employed to register and discovery the most pertinent IoT resource based on the user requirement.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Vandana C.P: Conceptualization, methodology, software, Validation, Writing- Original draft preparation, Visualization, Investigation Validation, Writing- Reviewing and Editing. Dr. Ajeet Chikkamannur: Supervision, Reviewing.

Acknowledgments

This research work is supported by R.L Jalappa Institute of Technology, Doddabalappur, Bangalore, India.

References

[1] X. Xiaojiang, W. Jianli, and L. Mingdong, “Services and key technologies of the internet of things”, *ZTE Communications*, Vol. 8, No. 2, pp. 26–29, 2020.

[2] CISCO, Internet of things. [Online]. Available: https://www.cisco.com/c/dam/en/us/products/collateral/se/internet-of-things/at-a-glance-c45-731471.pdf

[3] C. P. Vandana and A. A. Chikkamannur, “Study of Resource Discovery trends in Internet of Things”, *International Journal of Advanced*

Networking and Applications, Vol. 8,No. 3, pp. 3084-3089 ,2016

[4] C. P. Vandana and A. A. Chikkamannur, "S-COAP: Semantic Enrichment of COAP for Resource Discovery", *Springer Nature Computer Science SN COMPUT. SCI.* 1, No. 88, 2020.

[5] C. P. Vandana and A. A. Chikkamannur, "Semantic ontology based IoT-resource description", *International Journal of Advanced Networking and Applications*, Vol. 11, No. 6, pp. 3022–3030, 2019

[6] X. Xue and J. Chen, "Optimizing sensor ontology alignment through compact co-firefly algorithm", *Sensors*, Vol. 20, No. 7, pp. 2056, 2020.

[7] C. P. Vandana and A. A. Chikkamannur, "Semantic Annotation of IoT Resource with ontology orchestration", In: *Proc. of Third International Conf. on Advances in Electronics, Computers and Communications (ICAECC)*, Bengaluru, India, pp. 1-7,2020.

[8] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review", *Data Classification: Algorithms and Applications*, p. 37, 2014.

[9] Z. Zhao, L. Wang, H. Liu, and J. Ye, "On similarity preserving feature selection", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 25, No. 3, pp. 619–632, 2013.

[10] R. J. Urbanowicz, M. Meeker, W. L. Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: Introduction and review", *Journal of Biomedical Informatics*, Vol. 85, pp. 189-203, 2018

[11] C. P. Vandana and A. A. Chikkamannur, "Feature Selection: An Empirical Study", *International Journal of Engineering Trends and Technology*, Vol. 69, No. 2, pp 165-170,2021

[12] N. Hoque, D. Bhattacharyya, and J. K. Kalita, "MIFS-ND: A mutual information-based feature selection method", *Expert System Applications*, Vol. 41, No. 14, pp. 6371-6385, 2014.

[13] B. Venkatesh and J. Anuradha, "A hybrid feature selection approach for handling a high-dimensional data", *Springer Innovations in Computer Science and Engineering*, Vol. 74, pp. 365373,2019

[14] A. A. Aliane, H. Aliane, M. Ziane, and N. Bensaou, "A genetic algorithm feature selection based approach for Arabic Sentiment Classification", In: *Proc. of 13th IEEE International Conference of Computer Systems and Applications (AICCSA)*, Agadir, 2016, pp. 1-6

[15] C. Liu, W. Wang, Q. Zhao, X. Shen, and M. Konan, "A new feature selection method based on a validity index of feature subset", *Pattern Recognition Letters*, Vol. 92, pp. 1–8, 2017.

[16] S. Li and S. Oh, "Improving feature selection performance using pairwise pre-evaluation", *BMC Bioinformatics*, Vol. 17, No. 12, 2016

[17] X. Xue, M. Yao, and Z. Wu, "A novel ensemble-based wrapper method for feature selection using extreme learning machine and genetic algorithm", *Knowledge Information Systems*, Vol. 57, No. 2, pp. 389412, Nov. 2018.

[18] J. Zhang, Y. Xiong, and S. Min, "A new hybrid filter/wrapper algorithm for feature selection in classification", *Analytica Chimica Acta*, Vol. 1080, pp. 43-54, Nov. 2019.

[19] X. Zhu, Y. Wang, Y. Li, Y. Tan, G. Wang, and Q. Song, "A new unsupervised feature selection algorithm using similarity-based feature clustering", *Computational Intelligence*,Vol. 6, No. 3, pp. 1–21,2018

[20] Y. Meidan, M. Bohadana, A. Shabtai, J. D. Guarnizo, M. Ochoa, N. O. Tippenhauer, and Y. Elovici, "ProfilIoT: A Machine Learning Approach for IoT Device Identification Based on Network Traffic Analysis", In: *Proc. of the Symposium on Applied Computing (SAC)*, Morocco, pp. 506–509,2017

[21] S. Marchal, M. Miettinen, T. D. Nguyen, A. Sadeghi, and N. Asokan, "AuDI: Toward Autonomous IoT Device-Type Identification Using Periodic Communication", *IEEE Journal on Selected Areas in Communications*, Vol. 37, No. 6, pp. 1402–1412, 2019.

[22] B. Chakraborty, D. M Divakaran, I. Nevat, G. W. Peters, and M. Gurusamy, "Cost-aware Feature Selection for IoT Device Classification", *IEEE Internet of Things Journal*, doi: 10.1109/JIOT.2021.3051480

[23] K Clara, Suma, and V. Guruprasad, "Hybrid data model of PACE and quadruple: an efficient data model for cloud computing", *International Journal of Computer Aided Engineering and Technology*, Vol. 13, No. 12, pp.73–100, 2020.

[24] CoRE Resource Directory Link format-https://tools.ietf.org/html/draft-ietf-core-resource-directory-08

[25] Cooja Simulator – Contiki OS https://anrg.usc.edu/contiki/index.php/CoojaSimulator