CBARS: Cluster Based Classification for Activity Recognition Systems

Zahraa Said Abdallah[1,4], Mohamed Medhat Gaber[2], Bala Srinivasan[1], and Shonali Krishnaswamy[1,3]

[1] Centre for Distributed Systems and Software Engineering
Monash University,Melbourne, Australia
zahraa.said.abdallah@monash.edu,
srini@monash.edu
[2] School of Computing, University of Portsmouth
Portsmouth, Hampshire, England, PO1 3HE, UK
mohamed.gaber@port.ac.uk
[3] Institute for Infocomm Research (I2R), Singapore
shonali.krishnaswamy@monash.edu
[4] Faculty of computer and information Scienc
Ain Shams Univerity, Cairo, Egypt

Abstract Activity recognition focuses on inferring current user activities by leveraging sensory data available on today's sensor rich environment. Supervised learning has been applied pervasively for activity recognition. Typical activity recognition techniques process sensory data based on point-by-point approaches. In this paper, we propose a novel Cluster Based Classification for Activity Recognition Systems, *CBARS*. The novel approach processes activities as clusters to build a robust classification framework. *CBARS* integrates supervised, unsupervised and active learning and applies hybrid similarity measures technique for recognising activities. Extensive experimental results using real activity recognition dataset have evidenced that our new approach shows improved performance over other existing state-of-the-art learning methods.

Keywords: Activity recognition, Cluster based classification, Hybrid similarity measure

1 Introduction

There is a general consensus on the need for effective automatic recognition of user activities to enhance the ability of a pervasive system to properly recognise activities and react to circumstances. Recognizing human activities based on sensory data has recently drawn much research interest from the pervasive computing community. Activity recognition system focuses on inferring the current activities of users by leveraging the rich sensory environment. Sensor readings are collected and interpreted to recognise various human activities. Systems that can recognise human activities from sensory data opened the door to many important applications in the fields of healthcare, social networks, environmental monitoring, surveillance, emergency response and military missions.

Activity recognition (AR) is typically viewed as a classification problem where many traditional machine learning techniques can be applied [1]. In most existing supervised learning approaches in AR, the training data is collected, a classification model is generated offline from the collected data, and finally the obtained model is deployed to recognise the activity. A wide range of classification models has been used for activity recognition such as Decision Trees, Naive Bayes and Support Vector Machines. A typical activity recognition system builds the learning model with annotated data to recognise new data and predict human activity type based on the learning model.

We propose a novel cluster based classification method for robust activity recognition across users. We coined our technique *CBARS*, which stands for *Cluster Based Activity Recognition System*. *CBARS* adapts hybrid similarity measure classification for both accurate activity recognition and active learning for the new/unlabelled sensory data. Our proposed technique extends the state-of-the-art in AR by providing the following advantages.

- *Adaptability to the nature of activity recognition data:* People perform activities in a sequential manner (i.e., performing one activity after another). Therefore, activity recognition data stream typically composites of sequence of chunks that represents various activities. Different from other activity recognition systems, *CBARS* is a cluster based classification that deals with activities as clusters rather than processing each point. The novel approach is adapted for activity recognition data nature. Therefore, computation and processing time are conserved when dealing with the entire cluster instead of processing each point.
- *Hybrid similarity measure:* Learning model in *CBARS* contains clusters that represent different activities. When new cluster is emerged, hybrid similarity measure is deployed to match up similarities of the new cluster/activity with the existing ones. These measures are namely distance, density, gravity and within cluster standard deviation ($WICSD$). Applying the aforementioned similarity measures for activity recognition shows superiority over the use of individual ones, and therefore enhances the system robustness across users.
- *Combination of modelling techniques:* The system combines supervised, unsupervised and active learning all in one data stream model. We initially build the learning model with supervised learning. When new data received, unsupervised learning is deployed to cluster activities. Active learning is also employed in the event of confusion on cluster labels.
- *Framework for adapted model and evolving data stream:* One of the characteristics of *CBARS* framework is the flexibility to be updated as the data evolves. Therefore, the updated model is personalised and adapted to the most recent changes detected in the user's activities. In this paper, we present the framework that allows adaptation over time. However, the implementation of the adaptation is not in the scope of this paper.

To the best of our knowledge, no other existing activity recognition system addresses all aforementioned points in a single framework. The rest of the paper

is organised as follows. Section 2 provides a discussion of the research context. Explanation of the proposed framework and its details are presented in Section 3. Section 4 reports the experimental results and analysis. Finally, Section 6 concludes the paper with a summary.

2 Research Context

An efficient approach based on data mining has been recently proposed in a number of research projects considering the activity recognition from the machine learning perspective. Methods commonly used for activity classification were reviewed in [1]. Supervised learning has been deployed pervasively for activity recognition. One example system is explained in [2]. In this system, three classification techniques from WEKA [3] are used to induce models for predicting the user activities. Some other systems used fuzzy classifiers for activity recognition as in [4] and [5]. Few studies considered unsupervised learning techniques for activity recognition and change detection. In [7], the feasibility of applying a specific type of unsupervised learning to high-dimensional, heterogeneous sensory input was analysed. The correspondence between clustering output and classification input was proposed as well. Typically there is only a small set of labelled training data available in addition to a substantial amount of unlabelled training data. Therefore, some studies considered labelling only profitable samples of data or continue learning while system is running. Longstaff et al. Longstaff et al. [8] investigated methods of further training classifiers after a user begins to use them using active and semi-supervised learning.

To the best of our knowledge, none of these techniques has considered combining supervised, unsupervised and active learning for building a robust activity recognition system across users. Typical activity recognition stream is formed from a sequence of data chunks representing activities. Therefore, we propose a novel approach that treats data input as a stream and uses clustering to avoid having to respond to each input data point. As the stream evolves, there is a need to assess old and new clusters and this is handled with a hybrid similarity measure.

3 *CBARS*: Cluster Based Classification for Activity Recognition Systems

CBARS mainly composites of three consecutive phases. Illustration of the different phases is presented in this section.

3.1 Phase 1: Build Learning Model

Initially, supervised learning is applied on labelled data to train and generate the learning model. The generated model consists of set of clusters. Each cluster represents one of the labelled activities that applied while training the model.

CBARS creates k clusters of k activities in the training data. The cluster label is the majority label among cluster instances. Training examples typically contain outliers and noisy data that might affect the quality of the model directly. Therefore, we add a filtration step that aims to purify clusters and therefore build more accurate learning model.

Model Purification: While creating new cluster from training data, cluster is purified by considering only true-labelled instances inside the cluster and ignoring other mis-clustered examples (instances with different labels). Building on the purified clusters, characteristics are extracted for each cluster and all raw points are then dismissed.

CBARS extracts features from each cluster and dismisses all raw data at the end of this phase. Cluster characteristics are the basic information describes the cluster. Characteristics of a cluster include basically the cluster centroid, density, within cluster standard deviation and boundaries. The learning model is formed from set of clusters/activities that are deployed in training. The extracted features represent clusters/activities and raw data is dismissed.

3.2 Phase 2: Unsupervised Learning for New Data

This step aims to create clusters of various activities exist in data received. When unlabelled data emerged, we apply clustering on data to generate clusters of the performed activities . Various clustering techniques such as k-means, Expectation Maximisation and DBScan [9] have used and compared to reach the best performance. Clusters Characteristics are extracted and all raw data is dismissed similar to the learning model building procedure. The output of this phase is the set of clusters'/activities' characteristics ready for the recognition phase.

3.3 Phase 3: New Activity Recognition

As the stream evolves, we assess new and learning model clusters to predict new clusters' labels. This is handled with a hybrid similarity measure approach. As raw data has been dismissed, characteristics of the new cluster are compared to the existing learning model ones. The predicted label is based on the characteristics of the most similar cluster in the learning model. Clusters are similar if they match based on the hybrid similarity measure approach. For each new formed cluster, the algorithm checks how similar it is to other clusters already exist in the learning model. We apply various measures to test similarities among clusters. Each measure votes for its own " candidate" cluster from the measure respective. The predicted label is the candidate cluster with the majority of votes among all measures, while the true label of a cluster is the majority label among cluster instances. There are three cases expected from the voting procedure as follows.

1. ***Correct prediction:*** This case occurs when the majority of votes have chosen the candidate cluster/activity with the true label.

2. ***Active learning:*** In case of equal votes are assigned to a specific two clusters, user input is required to label the cluster. In this case, equal votes are assigned for exactly two clusters. The algorithm inquires about the correct label from user in an active learning mode with either of the labels of the two nominated clusters.
3. ***Incorrect prediction:*** The cluster is incorrectly classified when a total confusion with lowest confidence among all measures for candidate clusters or when voting for an incorrect cluster.

Hybrid similarity measure technique implements four measures namely distance, density, gravitational force and within cluster standard deviation. We concern in these measures about the distance among the cluster centroids, clusters density, how strong is the gravitational force among clusters and the cohesion inside the cluster. Learning model $\boldsymbol{LM}$ consists of $\boldsymbol{n}$ clusters/ activities.

$$LM = \{C_1, C_2, C_3, ...C_n\} \tag{3.1}$$

When new cluster C_{new} arrives, the algorithm deploys the various similarity measures to choose the best candidate cluster C_i from $\boldsymbol{n}$ clusters of LM. The four measures are:

- ***Distance:*** Cluster centroid is an n dimensional array of mean n- dimensional instances inside the cluster. C_i is the candidate cluster from distance perspective if distance between C_{new} and C_i centroids is the shortest among n clusters in LM.
- ***Density:*** Each cluster has its own density that distinguishes it from other clusters. Cluster density reflects the distribution of data points inside the cluster. It is described by the Formula 3.2 :

$$ClusDens = \frac{SizeOfCandidate}{AvgDist} \tag{3.2}$$

$$AvgDist = \frac{\sum_{i=1}^{m}(P_i - ClusCntr)}{m} \tag{3.3}$$

  Where (m) is the number of points in the cluster, (P) is a n- dimensional data point inside the cluster and $(ClusCntr)$ is the cluster centroid. the average Distance($avgDis$)is the within-cluster sum of the distances between cluster's examples and respective cluster centroid divided by the number of examples within cluster.Two clusters are similar from density perspective if they have the smallest difference in density. C_i is the candidate cluster if the difference between C_{new} and C_i density is the minimum among n clusters in LM.
- ***Gravitational force:*** Gravitational force has been previously applied in machine learning such as in [10] , [11], and [12] . There exists a natural attraction force between any two objects in the universe and this force is called gravitation force. According to Newton universal law of gravity, the strength of gravitation between two objects is in direct ratio to the product

of the masses of the two objects, but in inverse ratio to the square of distance between them. The law is described in Equation 3.4:

$$F_g = G\frac{m_1 m_2}{r^2} \tag{3.4}$$

Where F_g is the gravitation between two objects (clusters); G is the constant of universal gravitation; m_1 is the mass of object 1 (size of cluster 1); m_2 is the mass of object 2 (size of cluster 2); r the distance between the two objects (Euclidean distance between clusters' centroids) .
According to Equation 3.4, each cluster generates its own gravitational force created from its weight. The bigger the weight of the candidate the stronger the gravitational force produced around it. Therefore, the probability it could attract more data object would be increased. When the gravitational force between $Cnew$ and C_i is bigger than with other clusters existing in LM, then C_i is the candidate cluster from gravitational force perspective.

– ***WICSD (Within Cluster Standard Deviation):*** This measure considers the cohesion inside each cluster. Standard deviation of n dimensional points inside the cluster is calculated as the equation 3.5.

$$WICSD = \sqrt{\frac{\sum_{i=1}^{m} EDistance(P_i, ClusCntr)^2}{m}} \tag{3.5}$$

Where $EDistance$ is the Euclidean Distance, (m) is the number of points into the cluster, (P) is a n- dimensional data point inside the cluster and $(ClusCntr)$ is the cluster centroid.
Clusters that have similar standard deviation are more likely to present the same activity/label. C_i is the candidate cluster form $WICSD$ perspective if it has the smallest difference in standard deviation measure with C_{new} among n clusters in LM.

4 Experimental Study

This section reports the experiments conducted to study how $CBARS$ performs in practice. Activity recognition systems deal with high-dimensional, multi-modal streams of data. In the recently started European research project [13], the OPPORTUNITY dataset has been recorded to recognise complex activities. The dataset has labels for five users across five segments with annotated activities such as (sitting, walking and running) streaming form accelerometer sensors attached to the user's body.

Let F_c = total existing class instances correctly classified, F_a = total existing class instances trigger active learning, F_i = total existing class instances misclassified, S = total instances the dataset. We use the following performance metrics to evaluate our technique. M_c: % of class instances correctly classified = $\frac{F_c * 100}{S}$, M_a: % of class instances requires active learning = $\frac{F_a * 100}{S}$, and ERR: % of misclassification class instances = $\frac{F_i * 100}{S}$.

CBARS is tested on the OPPORTUNITY dataset. Part of the data is used to build the learning model; other new date is applied for testing. Testing data is a new data that has not used for training the model. Learning and testing data could be for the same user but different segments or for different users.

4.1 Cluster purification and learning model

Model purification is an essential step for pruning and refining the learning model. It has a superior advantage of building a robust model that filters outliers and wrong-labelled examples. Therefore, purified model represents activity of majority label with high confidence. Figures 1 and 2 show the effect of model purification on M_c with different runs N. Different runs are for different combinations of training and testing datasets for same user (N= 1,2) or various users (N= 3,4,5,6). As shown in figure 1, model purification shows better performance for most of the runs.

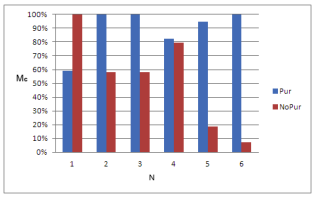


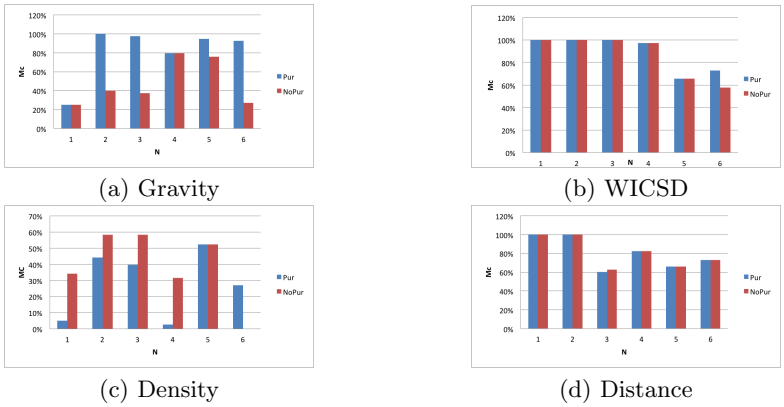
Figure 1: Cluster Purification and Recognition accuracy

The impact of the purification step on the various measures is shown in Figure 2. Eliminating outliers and wrong-labelled instances affects the position of clusters' centroids. Therefore, purification boosts measures that rely mainly on the coordinates of centroids. That include distance, gravity and WICSD as shown in Figure 2(a)(b)(d). On the other hand, purification might eliminate instances belong to low dense clusters that seemed to be outliers. Therefore, density might be affected badly with purification as showed in Figure 2(c). Although, purification has not enhanced the performance of all the measures, it has an overall positive effect on M_c when combining all of these measures as explained in Figure 1.

We evaluate *CBARS* performance with various clustering techniques deployed in phase 2. The algorithm implements Weka [3] clustering techniques, namely *k-means*, *EM* and *DBScan*. Experimentally, clustering accuracy has a direct influence on the system performance. Therefore, we apply the EM clustering in all our experiments unless otherwise stated.

4.2 Combining various measures

In the recognition phase, we apply a hybrid similarity measure technique for predicting new cluster's label. For the four similarity measures implemented in



(a) Gravity (b) WICSD

(c) Density (d) Distance

Figure 2: Cluster Purification and Measures accuracy

CBARS, each has its own average accuracy for correct prediction M_c of cluster's label. Distance measure has the best average accuracy of 68.48 % followed by the *WICSD* measure accuracy of 61.46%. Gravitational force and density measures come next with average accuracy of 54.17% and 53.13% respectively.

Combining the four measures benefits from the strengths of each one and eliminates encountered problems of using an individual measure and therefore helps enhancing the performance of single similarity measure techniques. Four different measure combinations showed the best performance. The four combinations are as follow. $Comb_1$ is for only the top accurate individual measures (distance, *WICSD*). Applying measures in $Comb_1$ attained a metrics of M_c= 79.76%, M_a= 20.24%. The recognition accuracy, M_c increased to 86.84%, while M_a decreased to 13.16% using $Comb_2$. This combination adds density to the before mentioned combination ($Comb_1$). $Comb_3$ includes gravity, distance and *WICSD*. Applying $Comb_3$ had the performance metrics of M_c= 85.96%, M_a= 10.21%. The highest correct recognition percentage attained when combining all measures in $Comb_4$. It has the recognition performance of M_c= 89.36% and M_a= 6.81%. As active learning percentage becomes higher, more frequent requests are sent to user to label confusing clusters. Therefore, the large percentage of active learning such as in $Comb_1$ and $Comb_2$ makes the system inefficient. Although *ERR* increases by 3.83% when combining all measures, this increase is not from the correct recognition rather than active learning percentage.

4.3 *CBARS* and other classification techniques

Table 1 illustrates the performance of *CBARS* across various learning and testing datasets in comparison to other iconic classifications techniques. Names of different datasets indicate the subject number such as *S1, S2, S3* followed by segment no such as *ADL1, ADL2, ADL3*.

Decision tree and the other classification techniques shows a good accuracy when the testing and training data are for the same user. However, testing data

Table 1: *CBARS* Recognition Performance

Train	***Test***	CBARS		DecisionTree	Naive Bayes	SVM	RFTree
		M_c	M_a				
$S1-ADL4$	$S3-ADL3$	**94.70%**	0.00%	34.78%	60.20%	47.83%	64.59%
$S1-ADL4$	$S1-ADL1$	**100.00%**	0.00%	79.57%	69.88%	82.47%	92.03%
$S1-ADL4$	$S2-ADL3$	68.56%	0.00%	37.86%	53.79%	70.70%	**74.32%**
$S1-ADL4$	$S1-ADL3$	**100.00%**	0.00%	88.17%	86.43%	90.31%	93.50%
$S1-ADL4$	$S2-ADL1$	**72.92%**	0.00%	44.51%	54.13%	58.80%	64.56%
$S3-ADL3$	$S1-ADL4$	53.19%	46.81%	42.41%	**83.53%**	66.42%	59.55%
$S3-ADL3$	$S2-ADL3$	68.56%	31.44%	48.36%	**82.24%**	58.87%	61.24%
$S3-ADL3$	$S2-ADL1$	**100.00%**	0.00%	51.76%	66.82%	56.36%	59.94%
$S3-ADL3$	$S1-ADL1$	**100.00%**	0.00%	39.79%	72.87%	65.09%	76.72%
$S1-ADL1$	$S1-ADL3$	59.17%	40.83%	83.90%	89.35%	90.94%	**92.66%**
$S1-ADL1$	$S2-ADL3$	**82.31%**	0.00%	53.86%	79.45%	72.93%	77.22%
$S1-ADL1$	$S3-ADL3$	**94.70%**	0.00%	40.27%	58.07%	51.70%	58.18%
Average		**79.75%**	15.64%	53.77%	71.40%	67.70 %	72.88%

across users confuses the decision tree and therefore has a negative impact on the recognition accuracy. On the other hand, *CBARS* shows stable high accuracy in recognising new activities either for the same or across users. The recognition accuracy (100%) in *CBARS* means that "all" new activities/clusters formed in testing phase have been successfully assigned a correct label. In case of active learning, *CBARS* is confused between two clusters labels . Therefore, user input is required to assign the true activity label. Figure 3 shows the percentage of incorrect prediction in *CBARS* and other classification techniques on various runs. Different runs are for different combinations of training and testing datasets for same user or across users. As shown in Figure 3, *CBARS* has the lowest *ERR* among all other techniques for all runs.

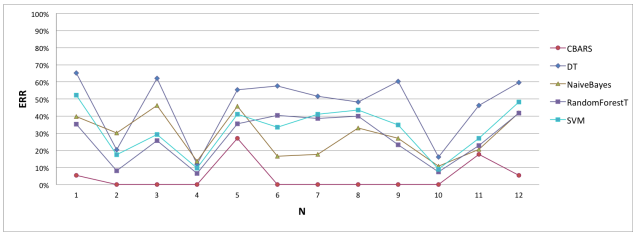


Figure 3: Incorrect Recognition for *CBARS* and other Classification techniques

5 Conclusion

In this paper we present a novel classification framework for activity recognition (*AR*) systems, named *CBARS*. This framework integrates supervised, unsupervised and active learning to build a robust and efficient recognition system. In comparison to state-of-art models, *CBARS* provides high performance results with the minimum error rate especially when dealing with recognition of activities across users.

References

1. Preece, S.J., Goulermas, J.Y., Kenney, L.P.J., Howard, D., Meijer, K., Crompton, R.: Activity identification using body-mounted sensors: a review of classification techniques. Physiological Measurement **30**(4) (2009) 1–33
2. Kwapisz, J.R., Weiss, G.M., Moore, S.A.: Activity recognition using cell phone accelerometers. SIGKDD Explor. Newsl. **12** (2011) 74–82
3. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. 2. edn. Morgan Kaufmann, San Francisco, CA (2005)
4. Helmi, M., Almodarresi, S.M.T.: Human activity recognition using a fuzzy inference system. In: IEEE International Conference on Fuzzy Systems. (2009) 1897–1902
5. Yang, J.Y., Chen, Y.P., Lee, G.Y., Liou, S.N., Wang, J.S.: Activity recognition using one triaxial accelerometer: A neuro-fuzzy classifier with feature reduction. In: Entertainment Computing - ICEC 2007, 6th International Conference. (2007) 395–400
6. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, B **39** (1977) 1–38
7. Li, F., Dustdar, S.: Incorporating unsupervised learning in activity recognition. In: Workshops at theAAAI Conference on Artificial Intelligence. (2011)
8. Longstaff, B., Reddy, S., Estrin, D.: Improving activity classification for health applications on mobile devices using active and semi supervised learning. In: Pervasive Computing Technologies for Healthcare (PervasiveHealth) 2010. (2010) 1–7
9. Ester, M., peter Kriegel, H., S, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise, AAAI Press (1996) 226–231
10. Peng, L., Yang, B., Chen, Y., Abraham, A.: Data gravitation based classification. Inf. Sci. **179**(6) (2009) 809–819
11. Abdallah, Z.S., Gaber, M.M.: Ddg-clustering : A novel technique for highly accurate results. In: Proceedings of the IADIS European Conference on Data Mining. (2009)
12. Abdallah, Z.S., Gaber, M.M.: Kb-cb-n classification: Towards unsupervised approach for supervised learning. In: Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2011. (2011) 283–290
13. Roggen, D., Förster, K., Calatroni, A., Holleczek, T., Fang, Y., Tröster, G., Lukowicz, P., Pirkl, G., Bannach, D., Kunze, K., Ferscha, A., Holzmann, C., Riener, A., Chavarriaga, R., del R. Millán, J.: Opportunity: Towards opportunistic activity and context recognition systems. In: Proc. 3rd IEEE WoWMoM Workshop on Autononomic and Opportunistic Communications. (2009)