Incorporating Unsupervised Learning in Activity Recognition

Fei Li and **Schahram Dustdar**
Distributed Systems Group
Vienna University of Technology
Argentinierstrasse 8/184-1
A-1040, Vienna, Austria

Abstract

Users are constantly involved in a multitude of activities in ever-changing context. Analyzing activities in context-rich environments has become a great challenge in context-awareness research. Traditional methods for activity recognition, such as classification, cannot cope with the variety and dynamicity of context and activities. In this paper, we propose an activity recognition approach that incorporates unsupervised learning. We analyze the feasibility of applying subspace clustering—a specific type of unsupervised learning—to high-dimensional, heterogeneous sensory input. Then we present the correspondence between clustering output and classification input. This approach has the potential to discover implicit, evolving activities, and can provide valuable assistance to traditional classification based methods.

As sensors become prevalent means in context detection and information channels proliferate to make context sharing easier, it is increasingly challenging to interpret context and analyze its effects on the activities (Lim and Dey 2010). We argue that applying traditional approaches to activity recognition may become more and more difficult to apply in context and activity-rich environments. In the literature, context attributes used for learning activities are chosen by either empirical assumption or dimension reduction to render a small set of features (Krause, Smailagic, and Siewiorek 2006). These approaches are infeasible in face of a broad spectrum of context information. The most significant drawback is that they fail to acknowledge the large variety of features needed to describe different activities.

For activity recognition, most previous works applied supervised learning approaches that aimed at predicting activities among a set of known classes(Ferscha et al. 2004). These approaches, however, are also challenged when coping with new and fast evolving activities. Unsupervised learning, particularly clustering, has been highly successful for revealing implicit relationships and regularities in large data sets. Intuitively, we can envisage an activity recognition approach that applies clustering to context history. The clusters, representing frequent context patterns, can suggest activities and their contextual conditions. The results can be used independently for analyzing and interpreting activities. Furthermore, clusters can reveal the scopes and conditions of activities and interactions. This information is valuable when determining the scopes of information sharing in pervasive environments.

Although clustering is a promising approach in discovering associations within context, the feasibility of traditional clustering is questionable in dealing with high dimensionality and heterogeneity of context data. In this paper, we will first conduct a detailed analysis about the challenges to apply clustering for activity recognition. Afterwards we introduce two recent subspace clustering methods that can address these challenges. Lastly, based on the analysis of unsupervised activity recognition, we will propose an activity recognition framework that incorporates clustering in conventional classification. We will show that the two directions are complementary to each other, and developing a hybrid approach will greatly benefit activity context awareness.

Unsupervised activity learning

Challenges

Given the various possibilities in collecting data, context contains a potentially large number of attributes. From a data analysis viewpoint, the variety and number of attributes is a double-edged sword. On the one hand, the more attributes we have, the more fine-grained associations we can get. On the other hand, more attributes can bring in more noise, obscuring hidden patterns in data. Classic clustering approaches, e.g. kNN (k-Nearest Neighbors), usually rely on various types of distance measures that are effective on low-dimensional data. However, on high-dimensional data the ”Curse of Dimensionality” becomes a significant problem because ”as dimensionality increases, the distance to the nearest data point approaches the distance to the farthest data point” (Beyer et al. 1999). Consequently, distance measures become ineffective. Dimension reduction techniques, such as Principle Component Analysis (PCA) (Wold 1987), can derive a single reduced set of dimensions for a whole data set, but at the same time they lose information about *local correlations of dimensions*. In activity recognition, an activity can be characterized by a certain set of dimensions, while another activity is characterized by a different set. Thus a single dimension reduction approach may not be suitable for activity recognition. We will illustrate the importance of local correlations through an example later.

Aimed at high dimensional data, subspace clustering or projected clustering (Kriegel, Kröger, and Zimek 2009; Parsons, Haque, and Liu 2004) has drawn considerable attention in the last years. The goal of subspace clustering is to find a list of clusters, each of a pair $< D, O >$, where D is a set of data attributes[1], and O is a set of data objects. In contrast to classic clustering methods, subspace clustering recognizes *local feature correlations* and finds similarities in data with regard to different subsets of dimensions. Subspace clustering have been applied to gene expression analysis, product recommendation, text analysis, etc.

Because activity recognition is highly dependent on discovering local correlations, subspace clustering is a promising direction in seeking an unsupervised activity recognition approach. However, there are more challenges for applying subspace clustering to context data. Most applications of subspace clustering use data consisting of homogeneous dimensions such as gene data. This is not the case in context data. Heterogeneity poses a strong limitation on choosing an effective clustering method. Similarity calculation, essential in many clustering approaches (Ester et al. 1996) and based on certain distance measures across multiple dimensions, is not applicable. Context data is heterogeneous in two respects:

- *Dimension semantics.* Each dimension in context data has unique semantics. The semantics decide the data type of each dimension, which can be numerical or nominal, continuous or discrete. It is not meaningful to measure the distance between two multidimensional points if the dimensions are semantically different.
- *Distribution.* Context data on each dimension is in a specific value domain and of a specific distribution. Although we can always define certain distance measure on each dimension, it is usually arbitrary and subjective to assign distance measures to non-numerical attributes. Even in a case of only numerical dimensions, different distributions of attributes can make normalization unreliable or even impossible, and eventually render similarity calculation defective.

A sample of context data is presented in Table 1. The sample contains a few fields to illustrate what is subspace cluster and how local correlations are reflected in clusters. Each dimension presents different semantics, has a different data type and value set. The data are an excerpt of a work day sensory records in a small company. B has to work from home that morning to wait for a plumbing service, and come to office in the afternoon. The data are collected every hour, and the dimension A.IMMsgNum means the number of IM messages A sent to the party in A.IMChat field in the last hour. Some clusters in one dimension can easily be observed. For example, in A's IMChat status, $\{o_0, o_1, o_2, o_7\}$ is a cluster indicating that he is mostly chatting with B. Some subspace clusters are also self-explanatory. The projection of objects $\{o_0, o_1, o_2\}$ on dimensions {A.Loc, B.Loc, A.IMChat, A.IMMsgNum} reveal that B is working from home and still closely in touch with A. Projecting $\{o_4, o_5\}$ onto {A.Loc, B.Loc}, we can understand A and B are having a meeting.

Applicable approaches

Based on previous analysis, we need a subspace clustering approach that can effectively handle heterogeneity of data. In other words, an approach that does not rely on distance calculation across more than one dimension. We found two possible candidates in the literature.

FIRES (FIlter REfinement Subspace clustering)(Kriegel et al. 2005) is a 3-step approach that avoids varying densities of different dimensionality in subspace clusters. The first step is *Preclustering*, which finds 1-dimensional (1D) clusters in all dimensions separately. Finding 1D clusters is easy since there is a rich selection of traditional clustering algorithms. Furthermore, we can choose different traditional clustering approaches for different dimensions according to their own patterns. With respect to context data, the significance of this step is that it treats attributes separately and transform them to a homogeneous structure–1D clusters, thus effectively solves heterogeneity problem. In the *generalization* step, the authors propose a highly scalable algorithm that merges 1D clusters in quadratic complexity w.r.t. the number of dimensionality. The third step, postprocessing, refines the subspace clusters found by generalization.

HSM (Heterogeneous Subspace Mining)(Mueller, Assent, and Seidl 2009) is a very recent algorithm dedicated to deal with heterogeneous data such as sensory inputs. Different to FIRES, HSM deals with heterogeneity while merging dimensions. The algorithm uses SCY-tree (Subspace Clusters with in-process-removal of redundancY) (Assent et al. 2008) as the supporting structure for dimension merging. There is no dedicated 1D clustering step in HSM. Each layer of SCY-tree consists of the data in one dimension. If the dimension is numerical, a grid-like merging is applied to find clusters in the dimension. The algorithm treats grids and nominal values the same while merging. To counter the problem of varying densities, the authors proposed a normalized density calculation approach that is applicable to subspace clusters containing both numerical and nominal dimensions.

A hybrid activity recognition process

The result of subspace clustering is a set of $< D, O >$ pairs, where D is a subset of all input dimensions $\mathbb{D}$, and O is a subset of all input data objects $\mathbb{O}$. Suppose $\mathbb{D}$ and $\mathbb{O}$ also represent the data space of training set for classification, users should label all objects in order to identify which class an object belongs to. The labeling process has some known limitations. Users can make mistakes when labeling due to wrong perception of situation. Labels provided by experts are not always up-to-date, and may not include new activities. On the other hand, large attribute set, i.e., high dimensionality of training data, also poses challenges to both classification algorithm and experts in data preparation. It is known that non-important attributes can obscure classification methods, and downgrade algorithm performance. Do-

[1] In this paper we use *attribute* and *dimension* interchangeably.

Object ID	Time	A.Loc	B.Loc	A.IMChat	A.IMMsgNum	A.Phone
o_0	09:30	office	home	B	23	occupied
o_1	10:30	office	home	B	26	available
o_2	11:30	office	home	B	18	available
o_3	12:30	out	in-car	inactive	0	available
o_4	13:30	meeting-room	meeting-room	inactive	0	available
o_5	14:30	meeting-room	meeting-room	inactive	0	available
o_6	15:30	office	office	C	12	occupied
o_7	16:30	office	meeting-room	B	5	occupied
o_8	17:30	out	out	inactive	0	available

Table 1: Sample context data

main experts may overlook some attributes or involve unnecessary ones when selecting attributes. In general, more sensory inputs make experts more prone to mistakes in selecting attributes.

Aimed at improving activity recognition capability in sensor and activity-rich environment, we propose a hybrid recognition process that utilizes the results of subspace clustering. The approach is illustrated in Figure 1.

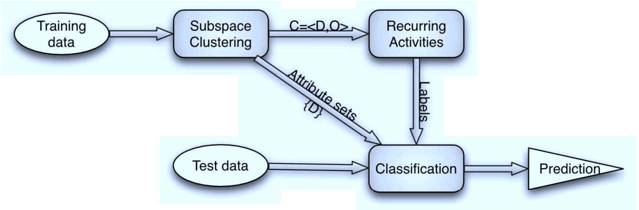

Figure 1: A hybrid activity recognition process

The output of subspace clustering is a set of clusters representing recurring activities $\{C\}$. Correspondingly, in classification the identified activities are used as labels. The subspace clusters can also identify different sets of associated attributes $\{D\}$, thus provide references for dimension reduction. The approach could benefit traditional activity recognition in several ways.

1. Providing initial labels. The found clusters suggest a set of identifiable activities, which can be directly named by experts rather than users. These names become initial labels. Although the list of clusters does not ensure the completeness of activities, the discovered label set can help reduce error rate in user labeling.
2. Finding unnoticed activities. Since unsupervised learning works with minimal assumptions about the result, it is possible to find implicit activities that may have been unnoticed. Periodically applying subspace clustering method can also help identify changes in the activities that users are often engaged with.
3. Identifying local attribute associations. This is the most significant advantage of subspace clustering. Among a large number of attributes, different activities may be associated with different attributes. The projected dimension set D is particularly useful when selecting attributes for classifier. For example, in Table 1, we can identify a close relationship between A's location (office) and his IM chatting party (B). We can even apply multiple classifiers to different associated attribute sets and in turn identify different groups of activities.
4. Finding overlapping activities. Since it is possible to have multiple projections of the same data object, overlapping of activities may be revealed in the original context history. In an environment of multiple activities conducted by multiple users, the overlapping is particularly valuable information that may suggest user groups, resource sharing, space sharing, parallelism of activities, etc. For example, projecting $\{o_0, o_6, o_7\}$ onto $\{$A.Loc, A.Phone$\}$, one can observe that A's phone usage is not tightly related to B's location, suggesting A's other activities in parallel with the continuous discussion with B.

Conclusion and further discussion

In this paper, we analyzed the feasibility and potential benefits of incorporating unsupervised learning methods in activity recognition. We have already started implementing the proposed approach and will carry out extensive experiments in the near future.

Although conceptually the proposed method is promising, there are still some further challenges to be addressed. In users' daily activities, there are some routine tasks that may lead to big clusters, e.g., A's phone status in Table 1—phone is available at most of the sampled time. Another extreme example is sensory input of emergency status, e.g. fire alarm, which in most cases will be standby. These statuses will naturally form big clusters. The significance of this type of big clusters is subject to specific situation, but at least in the previous examples, they do not add useful information to other subspace clusters. The approach to dealing with this type of attributes is still under investigation. The variety of sensors and activities challenge not only analytical methods, but also data fusion and activity context modeling. Activity context can be collected from diverse sources, including sensors for environmental information, personal mobile devices for individual status, status messages on social network, etc. Before applying context analysis methods, context fusion should first mediate, aggregate and process sensory data to ensure basic data quality. Context information should be presented by an exchangeable and extensible model that enables context sharing between machines. The model should also allow easy updates of activity description based on analytical results.

Acknowledgments

This work is supported by: 1) EU FP7 STREP Project SM4ALL (Smart hoMes for ALL) under Grant No. 224332, and 2) Pacific Control Systems LLC, Techno Park, Dubai, United Arab Emirates.

References

Assent, I.; Krieger, R.; Müller, E.; and Seidl, T. 2008. INSCY: Indexing Subspace Clusters with In-Process-Removal of Redundancy. In *2008 Eighth IEEE International Conference on Data Mining*, 719–724. IEEE.

Beyer, K.; Goldstein, J.; Ramakrishnan, R.; and Shaft, U. 1999. When Is “Nearest Neighbor” Meaningful? In Beeri, C., and Buneman, P., eds., *Database Theory — ICDT’99*, volume 1540 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg. 217–235.

Ester, M.; Kriegel, H.-P.; Sander, J.; and Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In Simoudis, E.; Han, J.; and Fayyad, U. M., eds., *Second International Conference on Knowledge Discovery and Data Mining*, 226–231. AAAI Press.

Ferscha, A.; Mattern, F.; Tapia, E.; Intille, S.; and Larson, K. 2004. Activity Recognition in the Home Using Simple and Ubiquitous Sensors. In Ferscha, A., and Mattern, F., eds., *Second International Conference on Pervasive Computing, PERVASIVE 2004*, volume 3001 of *Lecture Notes in Computer Science*, 158–175–175. Berlin, Heidelberg: Springer Berlin Heidelberg.

Krause, A.; Smailagic, A.; and Siewiorek, D. P. 2006. Context-Aware Mobile Computing: Learning Context-Dependent Personal Preferences from a Wearable Sensor Array. *IEEE Transactions on Mobile Computing* 5(2):113–127.

Kriegel, H.-P.; Kroger, P.; Renz, M.; and Wurst, S. 2005. A generic framework for efficient subspace clustering of high-dimensional data. In *Fifth IEEE International Conference on Data Mining*, 8 pp.

Kriegel, H.-P.; Kröger, P.; and Zimek, A. 2009. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Trans. Knowl. Discov. Data* 3(1):1–58.

Lim, B. Y., and Dey, A. K. 2010. *Toolkit to support intelligibility in context-aware applications*. New York, New York, USA: ACM Press.

Mueller, E.; Assent, I.; and Seidl, T. 2009. *HSM: Heterogeneous Subspace Mining in High Dimensional Data*, volume 5566 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg. 497–516.

Parsons, L.; Haque, E.; and Liu, H. 2004. Subspace clustering for high dimensional data: a review. *SIGKDD Explor. Newsl.* 6(1):90–105.

Wold, S. 1987. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems* 2(1-3):37–52.