

Attribute Constrained Rules for Partially Labeled Sequence Completion

Chad A. Williams^{1,*}, Peter C. Nelson¹, and Abolfazl (Kouros) Mohammadian²

¹ Dept. of Computer Science
University of Illinois at Chicago
851 South Morgan Street
Chicago, IL 60607-7053
cwilliam@cs.uic.edu, nelson@cs.uic.edu

² Dept. of Civil and Materials Engineering
University of Illinois at Chicago
842 West Taylor Street
Chicago, IL 60607-7023
kouros@uic.edu

Abstract. Sequential pattern and rule mining have been the focus of much research, however predicting missing sets of elements within a sequence remains a challenge. Recent work in survey design suggests that if these missing elements can be inferred with a higher degree of certainty, it could greatly reduce the time burden on survey participants. To address this problem and the more general problem of missing sensor data, we introduce a new form of constrained sequential rules that use attribute presence to better capture rule confidence in sequences with missing data than previous constraint based techniques. Specifically we examine the problem of given a partially labeled sequence of sets, how well can the missing attributes be inferred. Our study shows this technique significantly improves prediction robustness when even large amounts of data are missing compared to traditional techniques.

Keywords: frequent sequence mining, constrained rules, missing data.

1 Introduction

Frequent pattern mining of sequences has been a prominent research theme since its introduction by Agrawal and Srikant [1], yet how to effectively use pattern-based mining for classification and prediction remains a challenge [2]. The problem examined in this work is given a sequence of sets of attribute values where one to all fields within a set can be missing, populate the missing attribute values in the sequence. We refer to this problem as *partially labeled sequence completion*. Two common versions of this problem are:

* This research was supported in part by the National Science Foundation IGERT program under Grant DGE-0549489.

1. given the prior sequence, complete the missing elements of the current/next set; and
2. given a target set anywhere in a sequence complete the missing attribute values using the sets both before and after the target set

Many studies have focused on the next step prediction form, particularly for web applications such as pre-fetching and personalization [3,4]. In this study, we examine the more general form of the problem, since it also addresses the growing number of applications that would benefit from inferring additional information about an event given both the events before and after the event of interest. An example of this would be a group of mobile sensors that are periodically collected, where any number of the readings may be missing from any particular time step [5]. Other work in survey design suggests that if missing elements within a sequence can be inferred with a higher degree of certainty, it would greatly reduce the time burden on survey participants [6,7].

This work presents a new form of constrained sequential rules to address the more general form of the problem, which can also be applied for next step set prediction. Attribute constrained rules (ACR) are based on traditional sequential rules that can be derived from frequent sequential pattern mining, however extensions are made to better address attribute labeled sequences with missing value data. This problem of partially labeled sequence completion is formally stated below, followed by an overview of related work and an illustrative example of the value of ACR rules and how they are mined. Our study then shows this technique significantly improves prediction robustness for even large numbers of missing values compared to traditional sequential rules using a publicly available travel survey data set.

2 Partially Labeled Sequence Completion

Algorithms for rule mining of sequential patterns have been a major source of interest since they were first introduced in [8]. Much of prior work has focused on mining sequential rules for predicting future patterns, however being able to infer missing information within an observed sequence of sets of attribute values also has many useful applications that have largely been overlooked. One example of this is multi-day travel survey design, where being able to reliably infer missing values from the surrounding data set would allow respondent burden to be greatly reduced if only a portion of the data points needed to be collected regularly rather than the full set of questions [7]. To address problems such as these, we examine the more general problem of missing information within a sequence, however the technique also applies to traditional prediction as well. Specifically the problem we address is given a sequence where there are a known set of attributes that describe an event within the sequence, infer any missing values of the attributes for a target set. In this section, we formalize this constrained sequence problem and introduce a technique for mining and applying rules specifically for this task.

2.1 Problem Statement

For the problem of partially labeled sequence completion, let

$$H = \{H_1, H_2, \dots, H_n\}$$

be a database of sequences, and let:

$$H_i = \langle S_1, S_2, \dots, S_n \rangle$$

be the sequence of sets of observations in a sequence i ; where each observation set S_j is composed of 1 to m attributes $\{a_{j,1}, a_{j,2}, \dots, a_{j,m}\}$. Each attribute $a_{j,k}$ has a discrete set of values for the k^{th} position that is shared across all observation sets S . Intuitively the sequence H_i can be thought of as a series of recordings by a survey instrument or sensor with a fixed set of discrete measures (the attributes), where at each event j all measurements are relevant, but only a portion of these measures may actually be recorded in the set S_j . Given a sequence H_{target} of length l and a target set S_t to be completed where $1 \leq t \leq l$ and between 1 to m arbitrary attributes are missing values. Determine the values of all missing attributes $\{a_{t,1}, a_{t,2}, \dots, a_{t,m}\}$ in S_t . Thus our goal is to use the surrounding sequence information in H_{target} to populate any missing values to complete the set S_t .

3 Background and Motivation

3.1 Related Work

One of the common limitations of the majority of association mining algorithms is they do not take advantage of variable information or variable presence which becomes particularly important in sequences with missing values. For these types of sequences, which we refer to as partially labeled sequences, if we consider the sets being observed as a set of possible attribute assignments from a portion of the set of attributes (such as instrument output) we are observing, the problem of predicting future sets can become far more well defined in terms of the possible values given a set of possible attributes with either known or inferred constraints.

While techniques have been introduced for mining sequential patterns given regular expression constraints [9,10], the expression constraints in these works are best suited for matching a value pattern. For example, while an expression can be defined to match any sequence of values that can be described by a regular expression, the language does not provide for a more sophisticated notion of attribute value restrictions. While some aspects of this type of functionality could be encoded to restrict attribute values to a particular value such as the regular expression constraint: $\{“a_1 = 1”, (“a_2 = 2” | “a_2 = 3”)\}$, this type of encoding is insufficient for constraining the general presence of an attribute if all values are not known ahead of time.

Other work such as [11] has sought to identify the specific types of constraints needed for sequential pattern mining beyond those that can be expressed with

regular expressions. Their work introduces the concept of *item constraints*, which are intended to allow the presence (or absence) of a particular individual or group of items to be specified in the mining process. Given a particular query described in terms of constraints the algorithm they introduced, PrefixGrowth, finds patterns matching the constraints through a prefix extension method. While this algorithm is effective for finding patterns matching a particular query, it does not address being able to identify the set of possible constraint based rules for completing the values of a pattern in general.

3.2 Attribute Constrained Rule Mining

To address rules of this form we extend an idea from the text mining community called *label sequential rules* (LSR) [12,13]. Originally introduced for analyzing opinions in text reviews, this rule form was proposed for identifying common sentence forms or templates where a type of word of interest, termed a label, would likely appear. These rules form a more constrained matching pattern through wild cards producing rules of the form:

$$\langle \{1\} \{3, *, *\} \{6\} \rangle \rightarrow \langle \{1\} \{3, 4, 7\} \{6\} \rangle$$

where confidence of the rule would be defined with respect to the likelihood of the right hand side (RHS) given all sequences that contain the wild card constrained pattern. Thus, if we are only interested in rules that address completing two items in the middle set, these constraints allow a more meaningful measure of rule confidence since the likelihood is only measured in relation to patterns that match the LSR template.

For the task of partially labeled sequence completion, we propose a similar idea for identifying templates of sequential patterns of attribute values which we refer to as *attribute constrained rules* (ACR). Whereas with LSR the confidence of rules specify how likely a generalization is about elements within a pattern, with ACR the rule's confidence specifies the likelihood of a specific attribute value or combination of attribute values given a surrounding pattern.

Illustrative Example. In this section, we provide a set of illustrative examples of the benefit of constrained rules such as ACR and LSR. Throughout these examples refer to Table 1 as the sequence database.

Below we use the standard definitions of *support* and *confidence* defined as:

Definition 1. The *support* of the sequential rule $X \rightarrow Y$ is the fraction of sequences in the database that contain Y .

Table 1. Example sequence database

H_1	$\langle \{a_1\} \{a_2, b_2\} \{b_1\} \rangle$
H_2	$\langle \{a_1\} \{a_2, b_2\} \{a_2, b_1\} \rangle$
H_3	$\langle \{a_1\} \{b_2, c_2\} \{a_2\} \{b_1\} \rangle$
H_4	$\langle \{a_1\} \{a_2, c_1\} \{b_1\} \rangle$

Definition 2. The *confidence* of a sequential rule $X \rightarrow Y$ is the fraction of sequences in the database that contain X that also contain Y .

For an example of how constrained rules can better represent the applicable confidence, consider the following scenario: $H_{target} = \langle \{a_1, b_1\} \{a_2, ?\} \{b_1\} \rangle$, where $S_2 = \{a_2, \mathbf{b}:\?\}$ is the target set. The following traditional sequence associative rule would be applicable:

$$\langle \{a_1\} \{a_2\} \{b_1\} \rangle \rightarrow \langle \{a_1\} \{a_2, b_2\} \{b_1\} \rangle$$

[sup = 2/4, conf = 2/4]

Which can be interpreted as S_2 can be completed $\{a_2, b_2\}$ with a confidence of 2/4. By contrast a label constrained version of the same rule:

$$\langle \{a_1\} \{a_2, * \} \{b_1\} \rangle \rightarrow \langle \{a_1\} \{a_2, \mathbf{b}_2\} \{b_1\} \rangle$$

[sup = 2/4, **conf = 2/3**]

Where the notation $\{a_2, *\}$ indicates a set containing value a_2 for attribute a and a second value within the same set. As this example shows, by further constraining the attribute and location of pattern extension with LSR constraints, the confidence of the pattern is raised to 2/3 or roughly 67%. With ACR this idea is extended to constrain pattern matches to particular attribute values of interest. In our example, since we are specifically interested in the value of attribute b , the ACR version of the same rule would be:

$$\langle \{a_1\} \{a_2, \mathbf{b}:\ast\} \{b_1\} \rangle \rightarrow \langle \{a_1\} \{a_2, \mathbf{b}_2\} \{b_1\} \rangle$$

[sup = 2/4, **conf = 2/2**]

which is able to further clarify the confidence in populating attribute b , since it is able to discount sequence H_4 as it does not match the attribute constrained pattern. This advantage in accurately evaluating the value of the constrained sequence rule is the reason we examine ACR for partially labeled sequence completion.

The left hand side of Figure 1 shows the frequent sequence graph using a minimum support of 40% for Table 1 along with the support counts for each frequent sequence. All frequent ACR antecedents can be easily identified from the frequent item sets by expanding the remaining possibilities. For example the following antecedents can be generated from the frequent item set $\langle \{a_1\} \{b_2\} \{b_1\} \rangle$:

$$\langle \{a_1\} \{b_2\} \{b_1\} \rangle \Rightarrow \left[\begin{array}{l} \langle \{a : * \} \{b_2\} \{b_1\} \rangle, \langle \{a : * \} \{b_2\} \{b_1 : * \} \rangle, \\ \langle \{a_1\} \{b : * \} \{b_1\} \rangle, \langle \{a_1\} \{b : * \} \{b : * \} \rangle, \\ \langle \{a_1\} \{b_2\} \{b : * \} \rangle, \langle \{a : * \} \{b : * \} \{b_1\} \rangle, \\ \langle \{a : * \} \{b : * \} \{b : * \} \rangle \end{array} \right]$$

As this example shows, due to the combinatorial growth of the attribute constraint sets this problem quickly becomes impractical for datasets with a large number of attribute values or lengthy sequences if all completion rules are considered. For example with even this small example, the 17 frequent sequences

have over 40 potential ACR antecedents. For the problem as stated in Section 2.1 there are some properties that can be taken advantage of to reduce the number of ACR antecedents. Specifically, one of the key features of the problem we address is that every observation set S_j is composed of the same m attributes $\{a_{j,1}, a_{j,2}, \dots, a_{j,m}\}$, and only one target set is examined at a time. The implication of this fact is the property that for any set S_i , there is a value (whether known or not) for every attribute $a_{j,i}$. This property means that while the number of possible antecedents may grow quickly, the only ones that need to be kept are those with all constraints within a single set within the sequence. Once all possible ACR antecedents for the frequent pattern sets have been enumerated, the support for all the patterns can be updated with a single pass of the data set. By adding this subset and associated support to the frequent sequence graph as shown in the right hand side of Figure 1, ACR predictions and completion can be quickly determined using just the information in the extended frequent sequence graph. Note that while not shown in the figure, links would also exist between the ACR antecedents and the frequent sequence completions in the graph.

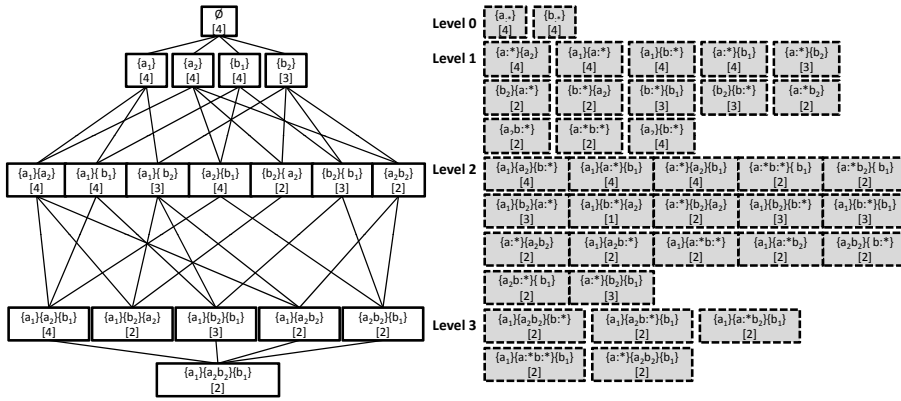


Fig. 1. ACR Frequent sequence graph

4 Experiments

4.1 Evaluation Metrics

For measuring prediction performance, we use the information retrieval metrics of precision and recall [14]. The basic definition of recall and precision can be written as:

$$\text{precision} = \frac{\# \text{ true positives}}{(\# \text{ true positives} + \# \text{ false positives})}$$

$$\text{recall} = \frac{\# \text{ true positives}}{(\# \text{ true positives} + \# \text{ false negatives})}$$

For the purpose of this study, since we are primarily interested in the correctness of the attribute value (if the attribute appeared at all). Thus *# true positives* is the number of attribute values predicted correctly; *# false positives* is the number of attribute values incorrectly predicted where the attribute did appear in the time step, and *# false negatives* is the number of attributes values which no prediction was made, but some value for the attribute appeared in the time step. Since these two measures are often associated with a tradeoff of one for the other, we also examine a combined metric the F-measure [15] which can be calculated as:

$$\text{F-measure} = \frac{(2 \cdot \text{precision} \cdot \text{recall})}{\text{precision} + \text{recall}}$$

We use this metric to compare the balanced performance of the algorithms.

4.2 Experimental Setup

Data. To evaluate the proposed ACR technique, we chose the 2001 Atlanta Household Travel Survey for several reasons. This dataset contains a large number of sequences of sets that are known to have a strong relationship between the entire set of attributes at each step and their ordering, making it well suited for sequential set prediction. Second, the type of data collected in this survey is very similar to one of the proposed applications of this type of partially labeled sequence learning, reducing survey participant burden [6,7]. Demonstrating that a significant portion of this type of data can be removed (i.e. a number of survey questions reduced) while limiting the impact on predictions is a significant step in showing the feasibility of this type of application. Finally, this data set represents one of the larger publically available data sets of this type, making the results of this study open to competitive comparisons by other work in this area.

The 2001 Atlanta Household Travel Survey was conducted from April 2001 through April 2002 on behalf of the Atlanta Regional Commission (ARC) [16]. The data consists of travel activity information for 21,323 persons from 8,069 households and 126,127 places visited during the 48-hour travel period. This survey focused on observed activity type, timing, and travel associated with each person's activity schedule captured over a 2 day period. The survey captured a wide range of in-home and out-of-home activity types which were broken down by a high-level classification. The survey captures over 250 attributes relating to the travel, activity, and demographic characteristics of the individual for each activity sequence that was recorded. The data is structured such that each event corresponds to an activity in the person's schedule with the set of attributes corresponding to the characteristics of the activity and travel associated with getting to the activity.

In the experiments below, we focus on a subset of 6 of these attributes: activity type, mode of transportation, arrival time, departure time, activity duration, and traveler age. These attributes were selected as they represent a mix of information about the type of activity, the travel, relative time the activity took place, activity duration, and features of the person involved that have been shown to

be highly related both intra-event and inter-event in predicting traveler activity patterns [17,18]. Thus, the dataset can be thought of as a database of sequences of events with the set of attribute values at each event being highly related. For the subset of the data set we use, there are 49,695 sets of activity information, with an average sequence length of just over 7.4 sets. Additional information about the data set can be found in Appendix A.

Methods. In the results below, we present the results of both the ACR technique, introduced in this work, and traditional sequence rules for a comparative baseline. As both rule-based techniques are frequent pattern based, which is deterministic for a given minimum support threshold, in all experiments below the same set of frequent patterns were used for both the traditional sequential mining and the ACR mining to ensure a fair comparison. In all experiments, both sets of rules were mined using the same minimum confidence threshold, and only rules that produced at least one target item in the target pattern were considered.

To generate predictions given a set of many potentially applicable rules, a ranking prediction scheme was utilized. Specifically, the rules were ranked in order by confidence, number of target productions, support, antecedent length, and finally a string based comparison to ensure repeatability if all other factors were equal. The productions of the top ranked rule were then applied to the target set, the remaining matching rules were then re-ranked as the number of target productions may have dropped due to the previous rule's productions. The rule with the highest rank of those remaining was then applied and this process continued until either a prediction had been made for all target items or no rules remained.

As described in Section 2.1, the problem of partially labeled set completion involves taking a sequence and trying to fill in or predict items within a single target set within a sequence. Since the problem of partially labeled set completion can take the form of predicting anywhere from a single item in a target set to all items in the target set, the results below reflect the average of all possible combinations of the target pattern in all possible positions for the target set. Where target pattern means: the set of attribute values in the target set that are being evaluated. Thus in the experiments below, for the target set any attribute value that is not specifically of interest as specified by the target pattern retains its original value for determining matching rules. For example if the target pattern included attributes a and c ($S_T = \{a_T c_T\}$). In testing the sequence:

$$\langle \{a_1 b_2 c_1\} \{a_2 b_2 c_2\} \{a_1 b_1 c_2\} \rangle$$

If the target set was S_2 for the sequence, the test sequence would thus be:

$$H_{target} = \langle \{a_1 b_2 c_1\} \{a_T b_2 c_T\} \{a_1 b_1 c_2\} \rangle$$

In the base experimental data set described above, no attribute values were missing. The missing data scenarios were created by randomly removing the specified percentage of values from both the training and test sets for any attribute appearing in the target pattern. All experiments below were run using a

minimum support threshold of 60% for frequent patterns and a minimum rule confidence threshold of 80%. To ensure the significance of our results, all results shown are the average of a 10 times cross-folding methodology.

4.3 Experimental Evaluation

ACR vs. Sequential Rule Comparison. In the first set of experiments, we examine the impact of missing data on each rule mining technique. Figure 2 portrays the recall of the missing attribute values. In the figure, *ACR-Avg* and *SEQ-Avg* represent the ACR results and the traditional sequential rules results respectively averaged over all possible number of target items in the target pattern. As these results show, the ACR technique produces both a higher overall recall as well as less degradation as values are removed and the data becomes sparser.



Fig. 2. Comparison of recall for ACR and traditional sequential rules as the percent of missing data increases

Since higher recall can indicate an increase in the quantity of predictions at the cost of accuracy, it is important to consider the precision as well. As Figure 3 shows, the boost in recall from ACR comes at a tradeoff of less than a .3% drop in precision. While this small drop is statistically significant, in practice the benefit of the additional 3-3.5% of attribute values with good predictions (recall) is likely to outweigh this small drop in precision for most applications.

The combined performance, as evaluated using the F-measure, is shown in Figure 4. As these results demonstrate, the ACR technique results in far better combined predictive robustness compared to traditional rules as the amount of missing data increases.

Number of Target Items Comparison. In the next set of experiments, a more in depth look is taken on the impact of the number of items trying to be predicted in the target pattern. In these results, *ACR-X* and *SEQ-X*

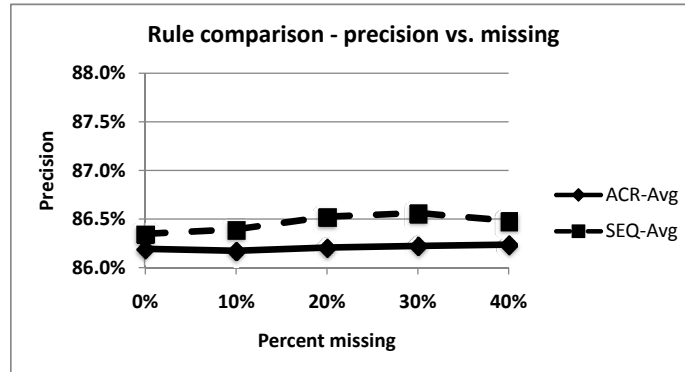


Fig. 3. Comparison of precision for ACR and traditional sequential rules as the percent of missing data increases

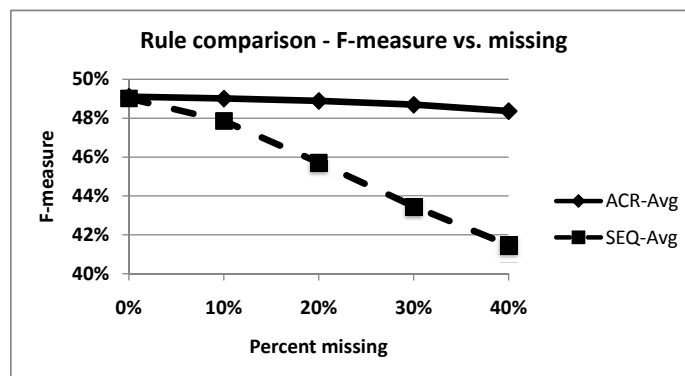


Fig. 4. Comparison of F-measure for ACR and traditional sequential rules as the percent of missing data increases

represent the ACR results and the traditional sequential rules results respectively averaged over all possible target patterns with X number of target items in the target pattern. Thus, $ACR-3$ would indicate the average performance of the ACR technique averaged across all possible target patterns with exactly 3 target items. The reason this is of interest, as Figure 5 shows, is that the number of items that are trying to be predicted can have a significant impact on how missing data affects prediction performance. As the results demonstrate for traditional sequential rules, while a small amount of missing data (10%) has a greater impact when predicting fewer items; as the amount of missing data increases (30%) this relationship is reversed.

As Figure 6 shows, with traditional sequential rules, as the number of values needing to be predicted increases, it becomes increasingly harder to recall all

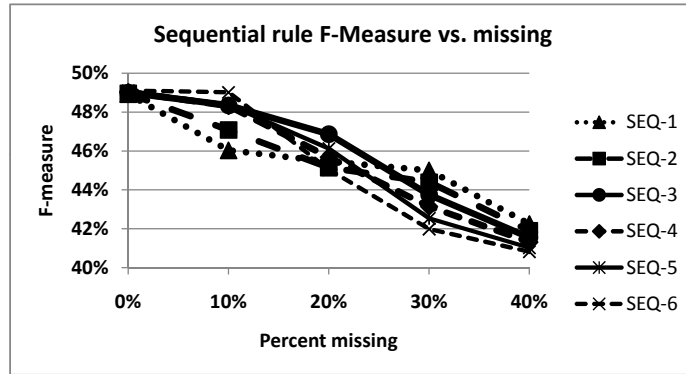


Fig. 5. Comparison of differences in predictive performance for traditional sequential rules for different target set sizes as the percent of missing data increases

of the target values and becomes even harder as the amount of missing data increases. Comparing this to the rules produced by the ACR technique shows that while this is still the case; the attribute constraints significantly increase the recall for this scenario even without missing values. Intuitively this is because in a less restricted target set, the attribute constraints better capture the likelihood of the predicted attributes all occurring simultaneously in the sequence than the unconstrained form of the rule.

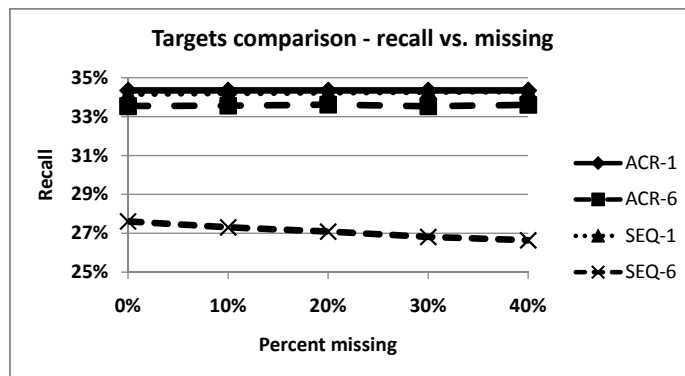


Fig. 6. Comparison of recall for ACR and traditional sequential rules for different target set sizes as the percent of missing data increases

A look at the precision for this same scenario, Figure 7, shows perhaps an intuitively unexpected result that the precision when predicting a full set is actually slightly higher than when predicting a single target item and furthermore increases slightly with some missing data. The reason for the higher precision with more target items is due largely to a smaller percentage of attribute values

actually being predicted (as reflected in the recall) and in this case, is likely in part due to a feature of the data set such that some attribute values are easier to predict than others (not shown in this work). Likewise the small elevation in precision associated with a percentage of the elements being removed likely reflects the benefit of randomly reducing the appearance of some of the less frequent items which may have some similarity to noise when considering full set prediction.

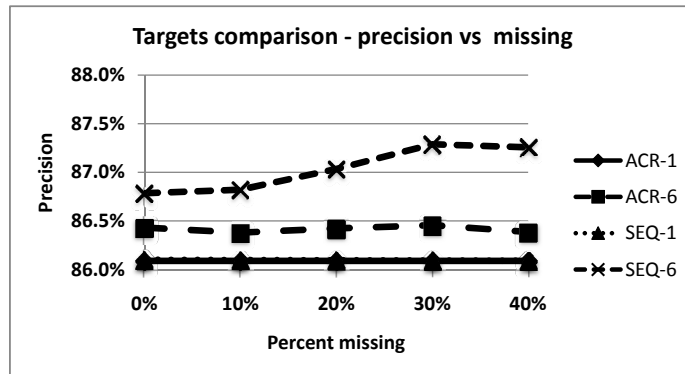


Fig. 7. Comparison of precision for ACR and traditional sequential rules for different target set sizes as the percent of missing data increases

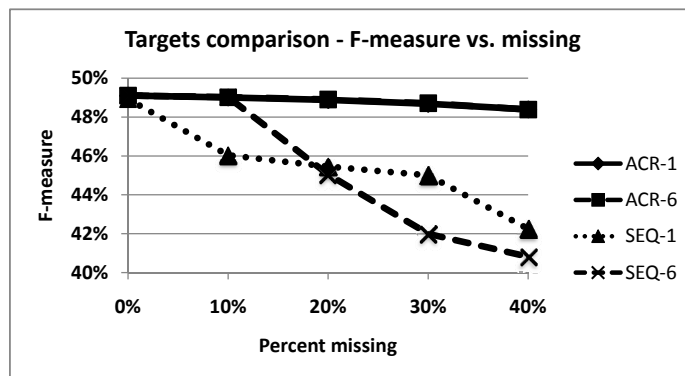


Fig. 8. Comparison of F-measure for ACR and traditional sequential rules for different target set sizes as the percent of missing data increases

Finally, the F-measure comparison of combined prediction performance is shown in Figure 8. As the results show, with traditional sequential rules while a small amount of data can be removed (less than 10% for this data set) with limited impact on full set prediction; as the amount of missing data increases beyond this the performance quickly degrades. Single target prediction displays a

slightly different trend being much more affected by even a slight increase in the amount of missing data, but being slightly more resilient than full set prediction as the amount of missing data increases beyond the initial amount.

This general pattern for traditional sequential rules show that the fewer the number of target items, the more significant any increase in missing data becomes; but also the less effected by subsequent increases in missing data is further illustrated in Figure 5. In contrast, the ACR technique proves much more resilient in either scenario as the amount of missing data increases, demonstrating nearly identical balance in predictive performance in both scenarios as the amount of missing data increases. This same nearly identical F-measure trend was observed for all target set sizes with the ACR technique (not shown).

5 Discussion and Future Work

In this work we introduced attribute constrained rules, a technique for mining and better estimating rule performance for partially labeled sequence completion. As our results demonstrate, this technique shows significant promise for accurately predicting sets of attributes within a sequence with missing data compared to a traditional sequential rule-based approach. In the context of survey applications aimed at reducing the time burden on participants such as those described in [6,7]; this represents a significant time savings opportunity. Specifically, the implications of the results presented in this work are that rather than needing to ask the full set of questions for each event as is currently done; a participant could be asked a much smaller portion of the questions with minimal impact on the benefits gained by pre-populating their likely responses. In the context of other applications such as mobile sensor data, this might represent a chance to reduce costly communication without reducing the ability to reliably predict missing values.

While the technique introduced is well suited for completing multiple attributes within a set of a sequence, a heuristic take on this technique may be necessary if it were to be used for predicting multiple sets simultaneously due to the combinatorial growth of possible ACR antecedents. In future work, we intend to explore ways this type of approach can be adapted to handle streaming data. Finally, a study is underway to confirm the benefits of this technique in practice for interactive survey applications such as those described above.

References

1. Agrawal, R., Srikant, R.: Mining sequential patterns. In: Yu, P.S., Chen, A.S.P. (eds.) Eleventh International Conference on Data Engineering, Taipei, Taiwan, pp. 3–14. IEEE Computer Society Press, Los Alamitos (1995)
2. Han, J., Cheng, H., Xin, D., Yan, X.: Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery* 15(1), 55–86 (2007)
3. Yang, Q., Zhang, H.H., Li, T.: Mining web logs for prediction models in www caching and prefetching. In: KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 473–478. ACM, New York (2001)

4. Mobasher, B., Dai, H., Luo, T., Nakagawa, M.: Using sequential and non-sequential patterns in predictive web usage mining tasks. In: *ICDM 2002: Proceedings of the 2002 IEEE International Conference on Data Mining*, Washington, DC, USA, p. 669. IEEE Computer Society Press, Los Alamitos (2002)
5. North, R., Richards, M., Cohen, J., Hoose, N., Hassard, J., Polak, J.: A mobile environmental sensing system to manage transportation and urban air quality. In: *IEEE International Symposium on Circuits and Systems, 2008. ISCAS 2008, May 2008*, pp. 1994–1997 (2008)
6. Marca, J.E., Rindt, C.R., McNally, M.G.: Collecting activity data from gps readings. Technical Report Paper UCI-ITS-AS-WP-02-3, Institute of Transportation Studies, Center for Activity Systems Analysis, University of California, Irvine (July 2002)
7. Auld, J., Williams, C.A., Mohammadian, A., Nelson, P.C.: An automated GPS-based prompted recall survey with learning algorithms. *Transportation Letters: The International Journal of Transportation Research* 1(1), 59–79 (2009)
8. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. *SIGMOD Rec.* 22(2), 207–216 (1993)
9. Garofalakis, M.N., Rastogi, R., Shim, K.: Spirit: Sequential pattern mining with regular expression constraints. In: *VLDB 1999: Proceedings of the 25th International Conference on Very Large Data Bases*, pp. 223–234. Morgan Kaufmann Publishers Inc., San Francisco (1999)
10. Garofalakis, M., Rastogi, R., Shim, K.: Mining sequential patterns with regular expression constraints. *IEEE Transactions on Knowledge and Data Engineering* 14(3), 530–552 (2002)
11. Pei, J., Han, J., Wang, W.: Mining sequential patterns with constraints in large databases. In: *CIKM 2002: Proceedings of the eleventh international conference on Information and knowledge management*, pp. 18–25. ACM, New York (2002)
12. Liu, B., Hu, M., Cheng, J.: Opinion observer: analyzing and comparing opinions on the web. In: *WWW 2005: Proceedings of the 14th international conference on World Wide Web*, pp. 342–351. ACM Press, New York (2005)
13. Liu, B.: *Web data mining: exploring hyperlinks, contents, and usage data*. In: *Data-Centric Systems and Applications*. Springer, Heidelberg (2007)
14. Cleverdon, C.: Evaluation of tests of information retrieval systems. *Journal of Documentation* 26, 55–67 (1970)
15. van Rijsbergen, C.: *Information Retrieval*. Butterworth, London (1979)
16. NuStats: 2001 atlanta household travel survey: Final report. Technical report, Atlanta Regional Commission (April 2003)
17. Timmermans, H. (ed.): *Progress in Activity-Based Analysis*. Elsevier, Oxford (2005)
18. Ettema, D., Schwanen, T., Timmermans, H.: The effect of location, mobility and socio-demographic factors on task and time allocation of households. *Transportation: Planning, Policy, Research, Practice* 34(1) (2007)

A Appendix: Data Set Information

The data selected for use in the experiments from the 2001 Atlanta Household Travel Survey was all events from the survey that didn't contain any missing values for the attributes: activity, mode of transportation, arrival time, departure time, duration, and age. This filtering was done to allow the experiments with

the amount of missing data conducted in this work to be performed in a more controlled manner. These attributes were broken down into the following discrete values:

- **Activity** - 29 values: 'Other at home activities', 'Working', 'Medical or dental', 'Getting ready', 'Major shopping', 'Personal', 'Watching children', 'Pick up something or drop off something', 'Worship or religious meeting', 'Visit friends or relatives', 'Household work or outdoors work', 'Entertainment', 'Outdoor recreation', 'Fitness or exercising', 'Rest or relax', 'Waiting for transportation', 'No other activities', 'Personal business', 'Other', 'Eating', 'Volunteer work', 'Work related from home', 'Community meetings', 'ATM, banking, post office, bill payment', 'Sleep', 'Work related business', 'School', 'Drop off or pickup someone', 'Incidental shopping'
- **Mode of transportation** - 12 values: 'Walk', 'Auto or van or truck - passenger', 'Other', 'Airplane', 'Intercity train (Amtrak)', 'Transit - MARTA bus', 'Dial a ride or paratransit', 'Intercity bus (greyhound, trailways)', 'School bus', 'Taxi, shuttle bus or limousine', 'Auto or van or truck - driver', 'Motorcycle or moped', 'Bicycle', 'Transit - CCT bus', 'Heavy rail - MARTA'
- **Arrival time** - 8 values: '3am-8am', '8am-10am', '10am-12pm', '12pm-2pm', '2pm-4pm', '4pm-6pm', '6pm-8pm', '8pm-3am'
- **Departure time** - 8 values: '3am-8am', '8am-10am', '10am-12pm', '12pm-2pm', '2pm-4pm', '4pm-6pm', '6pm-8pm', '8pm-3am'
- **Duration** - 7 values: '10 minutes or less', '10-30 minutes', '30-60 minutes', '1-2 hours', '2-4 hours', '4-8 hours', 'Greater than 8 hours'
- **Age** - 9 values: '10 years old or less', '10-20 years old', '20-30 years old', '30-40 years old', '40-50 years old', '50-60 years old', '60-70 years old', '70-80 years old', 'greater than 80 years old'