SPLICE-2 Comparative Evaluation: Electricity Pricing

Michael Harries
Artificial Intelligence Group
School of Computer Science and Engineering
The University of New South Wales, Sydney 2052, Australia.
E-mail: mbh@cse.unsw.edu.au

UNSW-CSE-TR-9905
July 1999

THE UNIVERSITY OF
NEW SOUTH WALES

School of Computer Science and Engineering
The University of New South Wales
Sydney 2052, Australia

Abstract

SPLICE-2 is a machine learning method designed for batch learning in domains with hidden changes in context. This report characterises the performance of SPLICE-2 on a real world dataset in comparison with C4.5, an on-line learner (emulated by C4.5), and an unsupervised learning system.

Two experiments are reported, both using electricity prices from the Australian state of New South Wales as the classification domain. The first experiment uses permutations of the dataset to explore the difficulties of using C4.5 to either identify hidden contexts, or to achieve the same accuracy as SPLICE-2 on this domain. The dataset permutations are also used to characterise some strengths and weaknesses of SPLICE-2. The second experiment uses permutations of an extended dataset from the same domain to examine, for both C4.5 and Splice-2, the effects of adding additional known attributes.

We find that C4.5 cannot induce the hidden contexts found by SPLICE-2. Further, SPLICE-2 generally provides more accurate results than C4.5. The exceptions occur when the order of the dataset is destroyed, or where new attributes are very similar to *time*. The best results from C4.5 are due to additional work on the part of the data analyst. One of the promises of SPLICE-2 is to reduce the level of additional work required of the analyst in such domains. We also find that a state-of-the-art conceptual clustering method does not identify the hidden context.

1 Introduction

SPLICE-2 [2] is a meta-learner designed for domains with hidden changes in context. SPLICE's job is to detect hidden contexts that are persistent over time[1], this includes, in particular, the detection of contexts that recur after alternation with other contexts. SPLICE applies a state-of-the-art propositional learner in the process of identifying each context and to produce a final description of each surmised contexts. Ideally, the output of SPLICE is as many decision procedures are there are contexts, which may then form a highly accurate combined classifier[2].

This work addresses the lack of existing work comparing the performance of SPLICE with existing state-of-the-art propositional learners on real world data.

2 Method

A batch learner, an on-line learner, and SPLICE-2 were applied to various datasets drawn from the electricity pricing domain. The goal was to compare accuracy, concept description size and success in detecting hidden changes in context.

The first experiment focused primarily on comparing the batch learner, C4.5, with SPLICE-2, in order to see if there were any circumstances under which C4.5 could detect large scale hidden contexts. To this end we tested various permutations of the dataset that might improve C4.5's chances of detecting these contexts. The on-line learner[3] is not directly equivalent to either C4.5 or SPLICE-2, as it does not provide a global description of the domain. Nevertheless, it provides a point of comparison between batch and on-line approaches to changing domains.

The second experiment used a richer dataset, permitting permutations of the dataset to be created with different numbers of relevant attributes. The learning systems were applied to each of three dataset permutations. In addition, a state-of-the-art unsupervised clustering system was applied to some of the datasets to confirm that hidden contexts were not available by unsupervised learning.

[1] SPLICE could also be used to detect hidden contexts persistent over some other attribute.

[2] The goal of having as many decision procedures are there are hidden contexts is likely to be confounded in the real world, as there can be many, overlapping hidden contexts affecting the same domain.

[3] The on-line learner is implemented by using C4.5 with a moving window.

2.1 Learning Systems

C4.5 [4] was used as our representative batch machine learning system. C4.5 was run with default options, with the exception that the subsetting option was turned on. The subsetting option permits C4.5 to construct branches of the decision tree based upon groups of more than one possible value of a nominal attribute. This option allows C4.5 to construct more succinct concepts than would otherwise be possible.

An on-line learner, designated *C4.5 on-line,* was implemented as a shell around C4.5 (again with default options plus subsetting). C4.5 was applied as an on-line learner by walking forward over the time ordered dataset. The process used follows:

> A concept was induced from a window containing the first k examples, and used to classify the r immediately following examples. Subsequent classification first moved the training window forward by r examples, induced a new classifier, then applied it to the r examples following the training window.

A variety of training window sizes, k, were tested, ranging from one week (336 examples), to ten weeks (3360 examples). A variable sized training window including all past examples was also tested. The test window, r, was kept at one week (336 examples) for this experiment.

SPLICE-2 [2] used C4.5 as the underlying learner (run with default parameters plus subsetting). A window size of 2000 items was used in order to focus SPLICE-2 on large scale contextual effects. The contextual clustering process was begun by randomly selecting a single instant of time upon which to split the dataset into two initial clusters. The contextual clustering algorithm was iterated 10 times to ensure convergence. (We provide an example of the impact of changing the number of contextual clustering iterations on both clusters found and accuracy for one of the following datasets.)

ACPro[4] is a commercial unsupervised clustering package available in MineSet (Silicon Graphics data mining package) based on the Minimum Message Length principle and is the commercial version of AutoClass [1]. This tool uses a combination of Bayesian statistics and the Estimation Maximisation algorithm to identify underlying clusters in data. It was used to cluster the data by attribute values for comparison with the contextual clusters found by SPLICE-2. These results are reported separately to those from the other learning methods.

Each learning system was applied to all datasets.

[4]Created by Wray Buntine of Ultimode (http://www.ultimode.com).

2.2 Metrics

The primary metric applied to assess each learner is accuracy. We also report on the size of the concepts learned as an indication of the quality of the concept. In addition, we show the hidden contexts identified by Splice-2.

2.2.1 Accuracy

For the learning methods C4.5 and Splice-2, a repeated 10 fold cross-validation was used to estimate the out of sample accuracy. f-fold cross-validation first randomly divides the dataset into f equal sized sub-sets, each of which is used in turn as a hold-out test set for classifiers trained on the remaining $f-1$ sub-sets. The resulting classification accuracies are averaged to give an overall estimated accuracy.

In ordinary cross-validation, the order of the training data is assumed to contain no information. Splice-2, however, relies upon the order of data examples to identify hidden contexts. Consequentially, we introduce a temporal cross-validation method in which the training set retains the original data ordering.

A further complication is that time ordered data can be temporally correlated. Two examples in close temporal proximity might well be very similar, or the same. If one of a pair of identical examples is allocated to the training set, and one to the test set, the resulting out of sample accuracy will provide an overly optimistic estimate of the real classifier accuracy.

To reduce this problem, we limit the likelihood of adjacent examples being assigned to both training and test sets. Rather than randomly distribute all examples, the data is first divided into periods, each containing k examples. These periods are then treated as indivisible units, which can be randomly assigned to one of the f cross-validation subsets. This limits the number of pairs of adjacent examples that can span both training and test sets. The following experiments used a k of 336 examples, or one week[5].

Yet another complication for cross-validating with Splice-2 is assessing accuracy on each held-out test set. As Splice-2 usually creates more than one stable concept, we must select a stable concept for each example in the test set. We used the following approach. For each example, x, in the test set, identify the example, y, from the training set, that occurred in the most recent past of x. The stable concept associated with example y is then used to classify example x.

To ensure that the reported results are accurate, the cross-validation process is repeated 10 times for each dataset and classifier. The results of

[5]This approach has some precedent. Kubat et al. [3] assess likely classifier accuracy by grouping data in known batches then running a type of restricted cross-validation, in which a single known batch is held out at each iteration.

these cross-validations were then averaged.

The reported accuracy of the on-line learner is an aggregate of the correctly classified items from a simulated on-line pass through the data.

2.2.2 Description

The concept size[6] is reported for C4.5 and SPLICE-2. Concept size in combination with concept accuracy gives an indication of the quality of the concept[7]. The concept size is calculated as the number of decision tree nodes used to describe the domain. For *C4.5* the reported description size is based on a single decision tree learned from the whole dataset. For SPLICE-2, the reported description size is an average over ten random initialisations of the algorithm. In addition, the SPLICE-2 description size is the total of the two stable concept sizes. We also draw a comparison between the individual stable concept sizes and the size of the *C4.5* decision tree.

Another aspect of domain description is the contextual clusters found by SPLICE-2; we show the clusters found by SPLICE-2 from ten random initialisations.

2.2.3 Significance

It is important to address issues of statistical significance. The types of data SPLICE-2 is designed to handle are particularly problematic for providing a measure of significance. Hidden changes in context imply that the data are sequentially time correlated, hence ordinary independence assumptions do not apply. We did, however, take care to run shuffled data, thus destroying the structure that SPLICE-2 is designed to find. The consistency of the degradation in accuracy induced by shuffling the data justifies a strong presumption that we are dealing with time correlated data, and thus cannot use standard significance measures.

2.3 Data - Electricity Prices

The data used in these experiments are based on the electricity market in the Australian state of New South Wales. This market was privatised on 11 May 1996. Prices in the market are set by matching, every five minutes, the current demand for electricity with the cheapest combination of electricity from all power stations according to price schedules published (and frequently updated) by each power station. Price schedules set the price for different levels of electricity production.

Electricity prices are affected by market demand and supply. Some major influences on market demand are season, weather, time of day, and central

[6]It makes little sense to report a concept size for the on-line learners, as the concept changes after every movement of the training window.

[7]By Occam's razor.

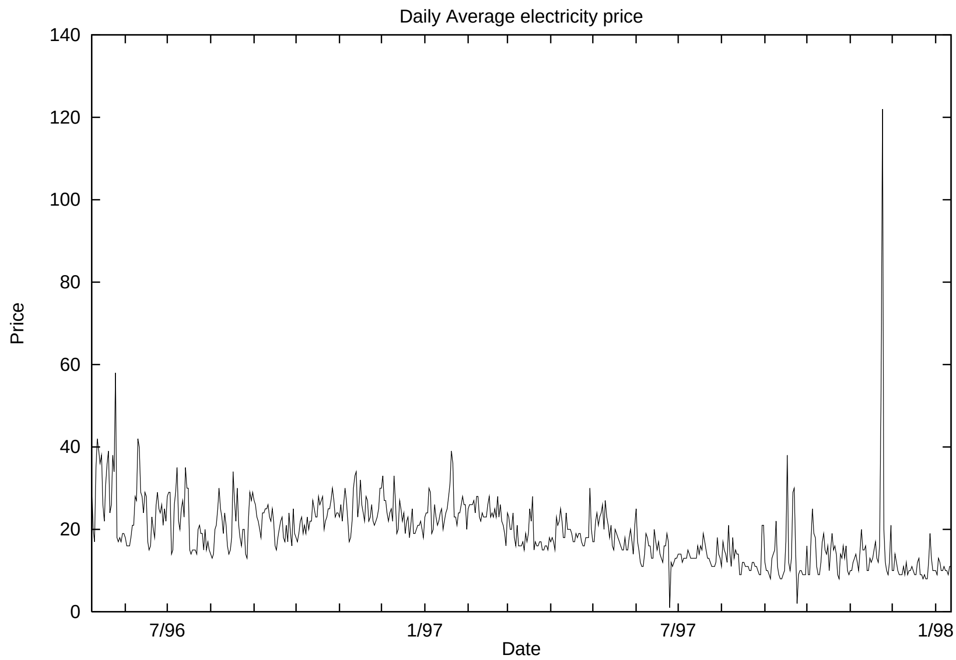

Figure 1: Daily average electricity price (based on the Elec1 dataset).

business district population density. The primary influence on supply is the number of electricity generators on-line. In other words, this domain is known to exhibit substantial seasonality and sensitivity to short-term events such as weather fluctuations. Figure 1 shows the average daily price over this period. As can be seen, the price level is not stationary. Figure 2 shows a two week fragment of the dataset at a finer resolution. Even at this level, quite dramatic changes in price behaviour can be seen day to day.

In addition to any recurring effects due to hidden changes in context, we can expect this domain to have evolved over time. As this dataset begins when the electricity market first opened, we can expect the behavior of participants to become more sophisticated over time, leading to changes in the aggregate market behaviour. In addition, on 4 May 1997, the electricity market in the neighbouring Australian state of Victoria was linked to the New South Wales market, allowing oversupply in one market to be sold interstate. This provided a dampener on some of the more extreme price movements.

Some effects seen in the dataset are due to unexpected conjunctions of circumstance. For instance, the large peak seen in Figure 1 at the end of November 1997 was caused by a break in the NSW/Victorian electricity connection coinciding with two very hot days in NSW.

Prior to this experiment, we had noted that SPLICE-2 could detect apparent seasonality in the first electricity dataset (description of Elec1-1 to

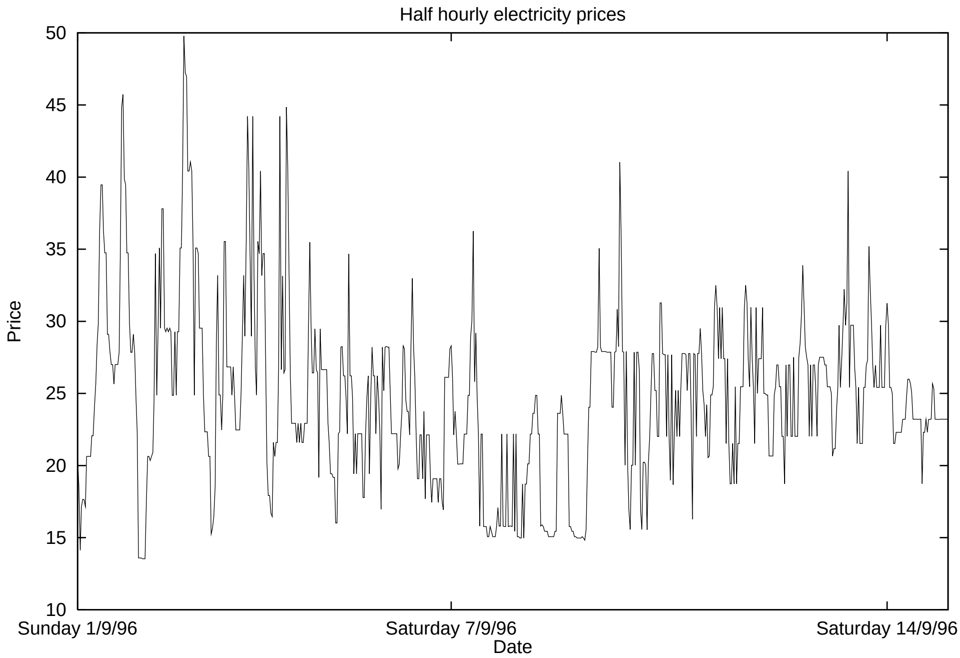

Figure 2: Electricity price, half hourly, two week sample.

follow). This is apparently due to the impact of gross weather patterns on power demand. Our domain expert[8] commented that, while this was certainly not explicit in the dataset, it does fit with their knowledge of the domain, and had not previously been automatically identified.

We use two datasets drawn from the electricity domain. The first remains proprietary and was provided by RMB Australia[9]. The second was drawn from the NSW electricity supplier directly. This data set runs to a later date and has additional attributes. The experiments on the first dataset (Elec1) were completed prior to gaining access to the second dataset (Elec2).

2.3.1 The Elec1 Dataset

The first dataset is based on 29,472 market snapshots from 11 May 1996 to 14 January 1998, each example of which is associated with a particular half hour.

Each example in the Elec1 dataset has a date (ignored), a day of the week (e.g. Mon), a time (e.g. 900), and a class label. The class label identifies whether the current price is higher (UP) or lower (DOWN) than a moving average over the last 24 hours (or 48 instances). As the class label only reflects deviations from the one day average price it removes the impact of

[8]Sandy Wilkinson of RMB Australia.

[9]Kim Horn originally suggested the use of this dataset for evaluating SPLICE.

Table 1: Fragment of the Elec1 dataset.

```
3/06/96, Mon, 2230, DOWN
3/06/96, Mon, 2300, DOWN
3/06/96, Mon, 2330, UP
3/06/96, Mon,  000, UP
4/06/96, Tue,  030, UP
4/06/96, Tue,  100, UP
4/06/96, Tue,  130, UP
4/06/96, Tue,  200, DOWN
4/06/96, Tue,  230, UP
4/06/96, Tue,  300, DOWN
4/06/96, Tue,  330, DOWN
4/06/96, Tue,  400, DOWN
4/06/96, Tue,  430, DOWN
```

Table 2: Datasets derived from Elec1.

Dataset	Description
Elec1-1	Elec1 - ordered by time (original dataset)
Elec1-2	Elec1-1 - augmented with an attribute for order
Elec1-3	Elec1-1 - augmented with an attribute for day length
Elec1-4	Elec1 - re-ordered by day length
Elec1-5	Elec1 - order shuffled

longer term price trends. The learning task is to construct a classifier that best identifies the class of previously unseen examples.

A fragment of the Elec1 dataset is shown in Table 1. This dataset has the following class distribution:

Classes		Default
UP	DOWN	Accuracy
12507	16965	57.6%

A number of permutations of this dataset, shown in Table 2, were explored.

Elec1-1 is based on the original data in the original, temporal, order. This is the dataset upon which SPLICE-2 had detected apparent seasonal hidden context prior to this experiment.

The following two dataset permutations, Elec1-2 and Elec1-3, were created to assist C4.5 in detecting, and taking advantage of any hidden context. Elec1-2 has an additional attribute showing the temporal order of each example; in effect a time-stamp. Elec1-3 has an additional attribute representing the day length, this was added to help C4.5 in finding the apparently *seasonal* hidden context identified by SPLICE-2. Note that the use of day length means that Spring and Autumn conflate.

Table 3: Elec2 attributes.

Date	ignored
Day of week	1..7
Time	based on half hour periods (1..48)
NSW electricity price	used to derive the target classes (ignored)
NSW electricity demand	numeric
Victorian electricity price	ignored
Victorian electricity demand	numeric
Schd. elec transfer between states	numeric
Class	NSW price higher/lower than 24 hour average

The next permutation, Elec1-4, re-orders the original data by day length in order to confirm that this attribute does reflect the hidden context. This allows us to both confirm that it was a valid choice of attribute in Elec1-3 and to test Splice-2 on an alternative data ordering that should also reflect context.

The final Elec1 permutation, Elec1-5, shuffles the original dataset in order to confirm that the information that Splice-2 is identifying is, in fact, implicit in the ordering of the data.

A difficulty for direct comparison between Splice-2 and C4.5 is that Splice-2 utilises information not available to C4.5. Some of the dataset permutations above have been selected to provide more direct comparisons. Two approaches were used: destroy the dataset ordering, thus reducing Splice-2 to the information available to C4.5; augment the dataset with an attribute for dataset order, thus allowing C4.5 access to the information used by Splice-2. The order-destruction approach was implemented in Elec1-5. The order-augmentation approach was implemented in the combinations: C4.5 on Elec1-2, and Splice-2 on Elec1-1.

2.3.2 The Elec2 Dataset

The second dataset contains 45,312 instances drawn from 7 May 1996 to 5 December 1998 with one instance for each half hour[10]. The attributes, in Table 2.3.2, include a number of actual demand figures. This is problematic as the actual demand would not be directly available ahead of time. However, we apply these attributes as a proxy for projections of attribute demand that might instead be used. These attributes should be very relevant to price as the price setting mechanism combines electricity demand with the price schedules published by power generators (supply).

[10]The Elec2 dataset was sourced from http://www.tg.nsw.gov.au. The dataset in C4.5 format is available from http://www.cse.unsw.edu.au/ mbh/context/elec.tar.gz.

Table 4: Fragment of the Elec2 dataset.

```
970626, 5, 41, 1460, 9232, 1423, 5626, 34, DOWN
970626, 5, 42, 1403, 8989, 1355, 5527, 274, DOWN
970626, 5, 43, 1385, 8630, 1321, 5379, 401, DOWN
970626, 5, 44, 1363, 8247, 1289, 5232, 407, DOWN
970626, 5, 45, 1403, 8491, 1317, 5123, 508, UP
970626, 5, 46, 1352, 8510, 1258, 5099, 531, DOWN
970626, 5, 47, 2371, 8371, 2212, 5431, 79, UP
970626, 5, 48, 1421, 8252, 1344, 5359, 240, UP
970627, 6, 1, 1423, 8105, 1335, 5260, 301, UP
970627, 6, 2, 1388, 7963, 1291, 5122, 439, DOWN
970627, 6, 3, 1413, 7750, 1326, 5293, 262, DOWN
970627, 6, 4, 1413, 7425, 1309, 5184, 324, DOWN
```

Table 5: Datasets derived from Elec2.

DataName	Description
Elec2-1	Elec2 - Using only Day of week and Time of day attributes
Elec2-2	Elec2-1 - Augmented with attribute for NSW electricity demand
Elec2-3	Elec2-2 - Augmented with attributes for Vic electricity demand and scheduled interstate electricity transfer

A fragment of the Elec2 dataset is shown in Table 4. This dataset has the following class distribution.

Classes		Default
UP	DOWN	Accuracy
19237	26075	57.5%

Table 5 shows the permutations of Elec2 used. Elec2-1 emulates the original attributes available in Elec1. This provides us with a comparison with the Elec1-1 results. Elec2-2 and Elec2-3 investigate the impact of supplying SPLICE-2 with attributes for electricity demand. In the case of Elec2-2 we add only the attribute for NSW electricity demand. In the case of Elec2-3 attributes for NSW demand, Victorian demand, and scheduled power transfer between states, are included.

3 Results

For brevity, this section provides a summary of the results. Full details are provided in the Appendix.

3.1 Elec1 Results

3.1.1 Basic Clustering Behaviour

To provide insight into the behaviour of SPLICE-2, we begin our results section with a demonstration of the impact of successive SPLICE-2 contextual clustering iterations. This serves to motivate the use of 10 SPLICE-2 clustering iterations throughout this article.

Figure 3 shows the clustering process for a single run of SPLICE-2 on the Elec1-1 dataset. Each chart in this figure shows the division of the domain into a number of clusters (Y-axis) over time (X-axis). The topmost chart shows the random partition initially selected by SPLICE-2. The next chart down shows the contextual clusters found after a single iteration of SPLICE-2 contextual clustering. Each chart thereafter shows the results of an additional SPLICE-2 clustering iteration. The bottom chart shows the final contextual clusters found by SPLICE-2. These are subjectively labelled as Summer, for the hot months, and Winter, for the cold months.

Figure 4 shows how the SPLICE-2 classification accuracy changes with sequential contextual clustering iterations. This figure shows the accuracy achieved (Y-Axis) against the number of clustering iterations (X-axis). The initial iterations can be seen to give the majority of improvement in accuracy (2.9% for the first five iterations); subsequent iterations give rather less improvement (0.1% for the last five iterations). These results are drawn from the Elec1-1 dataset, and are based upon 10 repetitions of a ten-fold cross-validation for each point. Although the dataset apparently converges to peak accuracy in the first five iterations, converging on stable clusters can take longer, as seen in Figure 3. Accordingly, we used the 10 SPLICE-2 contextual clustering iterations to ensure convergence.

3.1.2 Elec1 Comparisons

Table 6 shows the results achieved by C4.5 C4.5 on-line and SPLICE-2 on the Elec1 datasets. The detailed versions of these results are in Sections A.1, A.2, A.3, A.4, and A.5. For C4.5 both the average accuracy and average concept size in nodes are shown. For *C4.5 on-line*, only the accuracy is shown. For SPLICE-2, the accuracy, and total concept size in nodes are shown. In addition, SPLICE-2 has a column for the stability of clustering, this gives an indication of how frequently the same set of clusters was found in ten random initialisations on the same dataset, this figure is based on a subjective assessment of the contextual clusters found. The stability column may be read as follows: 2/10 means that of ten SPLICE-2 runs, two resulted in similar sets of contextual clusters; $(3 + 2)/10$ means that there were two groups of similar clusters, one having three similar sets, and another with two similar sets. The stability result for Elec1-5 contains an italicized *2*, this indicates that two of the SPLICE-2 runs converged to a single cluster.

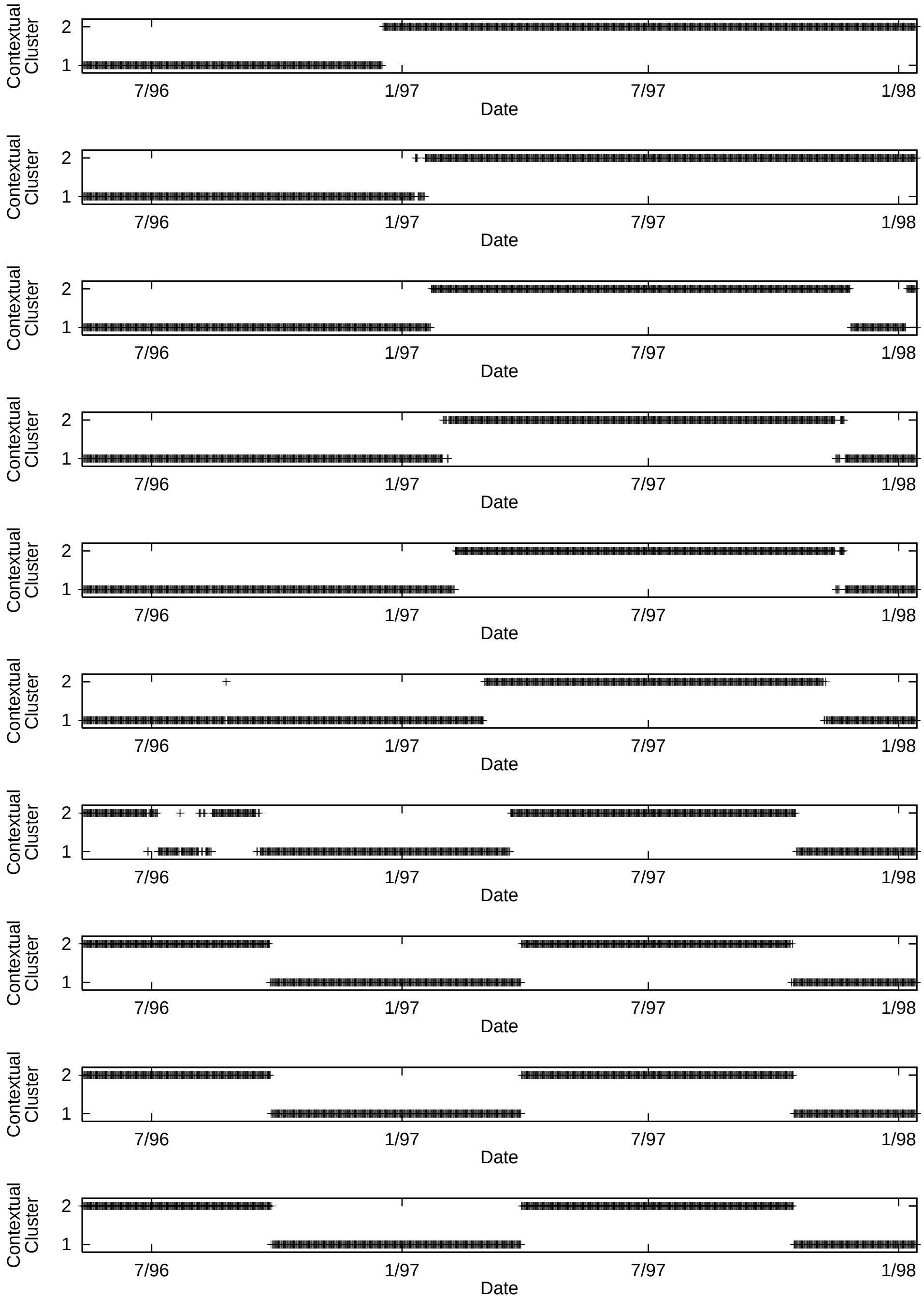

Figure 3: Splice-2 clustering on Elec1-1. This figure shows the impact of each iteration of the contextual clustering algorithm. At top is a random partitioning of the data, at bottom is the final clustering result.

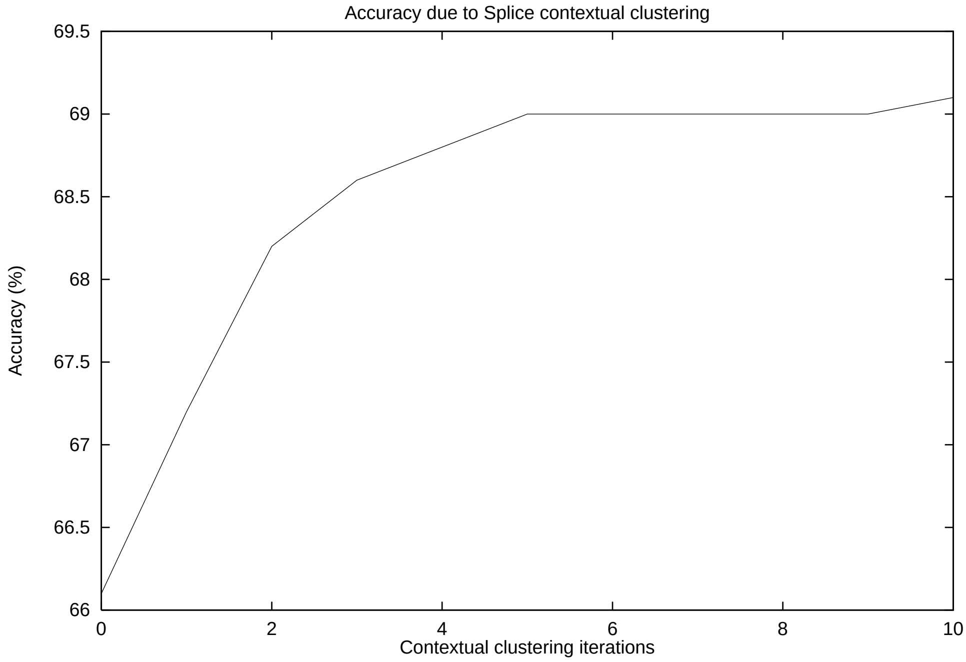

Figure 4: Accuracy at successive SPLICE-2 contextual clustering iterations

Table 6: Combined results for the Elec1 datasets.

Dataset	C4.5		C4.5 on-line	SPLICE-2		
	Accuracy	Nodes	Best Acc.	Accuracy	Nodes	Stable?
Elec1-1	65.7%	41	68.4%	69.1%	93.2	10/10
Elec1-2	68.1%	1535	65.2%	67.7%	1924.2	2/10
Elec1-3	66.1%	1045	65.4%	65.0%	1767.2	(3+2)/10
Elec1-4	65.7%	41	68.3%	69.4%	85.0	10/10
Elec1-5	66.3%	41	65.9%	66.1%	77.4	(2+3+2)/10

On Elec1-1, the original dataset, SPLICE-2 improved on the accuracy achieved by C4.5 with a difference in accuracy of 3.4%. SPLICE-2 also correctly classified 0.7% more original examples than *C4.5 on-line*. The SPLICE-2 clustering was consistent 10 times out of 10, (Figure 8) and follows the Summer/Winter pattern seen in the last chart of Figure 3.

Comparing the accuracy for C4.5 on Elec1-1 (65.7% from Table 6) and the first point in Figure 4 is interesting, as it appears that an improvement of 0.4% accuracy was be due to randomly partitioning the dataset without any contextual clustering. This is not surprising given a different mix of contexts in each subset, as each classifier can *specialise*, and thus improve the combined classifier accuracy.

On Elec1-2, C4.5 did quite well, with an accuracy of 68.1%. At the same time, the addition of an order attribute reduced the best C4.5 on-line result to 67.7% On this dataset, C4.5 beat SPLICE-2, this was probably due to SPLICE-2 clustering being confounded by an attribute that can be used to uniquely identify all periods of the dataset. Only two runs out of 10 led to similar clusters and in no case were there any recurring contextual clusters (Figure 19).

On Elec1-3, C4.5 again out-classified SPLICE-2 but with a difference of 1.1%. *C4.5 on-line* beat SPLICE-2 by 0.4%. SPLICE-2 clustering was more stable, but again did not induce any recurring contextual clusters (Figure 20).

On Elec1-4, we returned to the original dataset, with the data now ordered by day length. Here, SPLICE-2 had a 3.7% edge over C4.5. SPLICE-2 induced highly stable clusters (Figure 21). *C4.5 on-line* achieved a higher accuracy than any of the C4.5 accuracies.

On Elec1-5, the original dataset was shuffled, in order to destroy order information. Here, the accuracy results for all three learners were very similar, with C4.5 beating SPLICE-2 by only 0.2% and SPLICE-2 beating *C4.5 on-line* by only 0.2%. This result accords with the expectation that SPLICE-2 exploits information embedded in the order of the dataset. Without this information, SPLICE-2 can do no better than C4.5. This result demonstrates that the data are not independent and identically distributed and thus that significance measures should not be used. It is interesting that SPLICE-2 found some structure in the shuffled dataset (Figure 22). This is probably because the data was shuffled only once and contains spurious structure. In addition, two of the runs converged to a single cluster. This is what we might ideally hope to see from a randomised dataset.

3.2 Elec2 Results

Table 7 shows the combined results on the Elec2-1, Elec2-2, Elec2-3 datasets. The columns are the same as for Table 6. The table summarises detailed results in Sections A.6, A.7, and A.8.

Table 7: Combined results for the Elec2 datasets.

Dataset	C4.5		C4.5 on-line	SPLICE-2		
	Accuracy	Nodes	Best Acc.	Accuracy	Nodes	Stable?
Elec2-1	65.5%	54	68.4%	67.3%	130	10/10
Elec2-2	67.2%	386	68.4%	67.6%	536.1	(4+3)/10
Elec2-3	67.5%	1221	67.7%	67.6%	1607.7	(3+3+2)/10

With the extended basic dataset, Elec2-1, SPLICE-2 correctly classifies 1.8% more of the dataset than does C4.5. This is a little disappointing given the larger difference between SPLICE-2 and C4.5 on Elec1 and indicates that there may be are additional contexts or random concept drift represented in this, *larger*, data set. This conclusion, of additional embedded context change, is backed up by *C4.5 on-line* consistently outperforming both C4.5 and SPLICE-2. In addition, the clusters found by SPLICE-2 on Elec2-1 are stable in 8/10 cases but have two variants for the final context change (Figure 23), indicating some contextual ambiguity. In Elec2-2, an attribute for NSW electricity demand was added to the dataset, this marginally improved SPLICE-2 accuracy, and substantially improved C4.5 accuracy, although SPLICE-2 was still slightly more accurate with a difference of 0.4%. In both cases, there was a substantial increase in the classifier size. The clusters found by SPLICE-2 are interesting, as despite the reduction in accuracy, there were some stable clusters with recurrence (Figure 24).

In Elec2-3, attributes for NSW electricity demand, Victorian electricity demand, and the scheduled interstate electricity transfer, were added. This had no effect on the accuracy by SPLICE-2, but improved the C4.5 accuracy to only 0.1% less than SPLICE-2. In both cases, there was a substantial increase in the number of nodes. The SPLICE-2 clusters had some similarities in timing for context changes, and identified no recurring contextual clusters (Figure 25).

We conjecture that one of the reasons for SPLICE-2 doing less well on this dataset was that more than two contexts were represented. Table 8 shows the results for running SPLICE-2 on these data with three contextual clusters. (Based on Sections A.9, A.10, and A.11.) SPLICE-2, with three clusters, was able to equal *C4.5 on-line* for Elec2-1, almost equal it for Elec2-2, and beat it for Elec2-3. Even in the case where *C4.5 on-line* was more accurate, the result is good as it was achieved with a much smaller description cost than that effectively achieved by *C4.5 on-line* (i.e. 133 weeks times an estimated short term classifier size of at least 40 nodes: a minimum of 5320 nodes!). It is also worth noting (Figure 26) that SPLICE-2 frequently identifies a shoulder season when applied with three clusters.

Table 8: SPLICE-2 results for the Elec2 datasets given three contextual clusters.

Dataset	SPLICE-2		
	Accuracy	Nodes	Stable?
Elec2-1	68.4%	130	8/10
Elec2-2	68.2%	616	1/10
Elec2-3	67.9%	1818	1/10

3.2.1 Unsupervised Clustering

This section reports upon an application of the unsupervised learning system ACPro to the electricity datasets. The goal of this sub-experiment was to confirm that the contextual clusters identified by SPLICE-2 differ from the types of *conceptual* clusters that can be found by unsupervised clustering. ACPro is a commercial successor to the state-of-the-art unsupervised clustering system AutoClass [1]. This experiment confirms that contextual clusters are not easily induced by a state-of-the-art unsupervised learning system.

ACPro was applied the dataset permutations, Elec1-1, Elec2-1, Elec2-2, and Elec2-3. A variety of configuration options were explored.

Elec1-1 The Elec1-1 permutation of Elec1 has only the original Elec1 attributes, *day of week*, and *time of day*. Both of these attributes are known to be equally distributed, so do not provide much information for unsupervised learning. *Day of week* is cyclic and repeatedly runs through Monday to Sunday; equally, *Time of day* is cyclic and runs from 0:00 to 23:30 in 30 minute increments. Accordingly, there should be no irregularities for a clustering method to exploit.

There were three clustering attempts made on Elec1-1. In each, *day of week* was treated as a discrete attribute. In the first two attempts, *time of day* was treated as a continuous attribute. In the second attempt, ACPro was provided with additional processing duration; 3600 seconds rather than the default of 1800. In the third attempt, *time of day* was treated as a discrete attribute.

In attempt one, two clusters were found after 19 clustering cycles. These clusters differ only in the average *time of day*. One cluster had an average *time of day* of 5:00 am and the other had an average *time of day* of 4:45pm. In attempt two, three clusters were found after 39 clustering cycles. Again these clusters differ only in the *time of day*, with one cluster having an average *time of day* of 2:06am, another having an average *time of day* of 7:00pm, and the last with an average *time of day* of 9:33am. In run three, only a single cluster was found after 7 clustering cycles, containing all items.

It is interesting that ACPro found any clusters in this data, given its known structure. This shows that, under some circumstances, ACPro can produce *spurious* clusters.

Elec2-1 The Elec2-1 permutation of Elec2 has only the two basic attributes *day of week* and *time of day.* Again, there should be no irregularities in the data to be exploited by a clustering algorithm.

Three clustering attempts were made on Elec2-1. In each of them, day of week was treated as a discrete attribute. In the first two attempts, time was treated as a continuous attribute. In the first attempt, ACPro was run for the default duration of 1800 seconds; in the second, it was run for 3600 seconds. In the third attempt, both time and day of week were treated as continuous attributes.

In attempt one, two clusters were found after 22 clustering cycles. These clusters differ only in time of day. For this dataset, *time of day* is represented by the number of half hour periods since midnight. One cluster had an average number of periods of 11.8 (5:54am) and the other an average number of periods of 34.7 (5:21pm) In attempt two, two clusters were again found, this time after 31 clustering cycles. These clusters were based on average times of 10.4 periods (5:12am) and 33.2 periods (4:36pm).

In attempt three, ten clusters were found! These clusters also used the day of the week. Figure 5 show the cluster distribution over a week.

Elec2-2 One clustering attempt was made on Elec2-2. *Day of week* was treated as a nominal attribute and both period and NSW demand were treated as continuous attributes. The result was 27 clusters found after 85 clustering cycles. Figure 6 shows the clusters distributed over time.

Elec2-3 One clustering attempt was made on Elec2-3. *Day of week* was treated as a nominal attribute and *period*, *NSW demand*, *Victorian demand*, and *scheduled transfer* were treated as continuous attributes. The result was 50 clusters found after 50 clustering cycles. Figure 7 shows the clusters distributed over time. This is a much more interesting set of clusters in some ways, as it clearly shows a break at the point where interstate electricity transfers began, and also shows clusters with some larger scale temporal structure.

These results serve to show that the nature of clusters found by unsupervised conceptual clustering on these data is quite different to the types of contextual clusters identified by SPLICE-2.

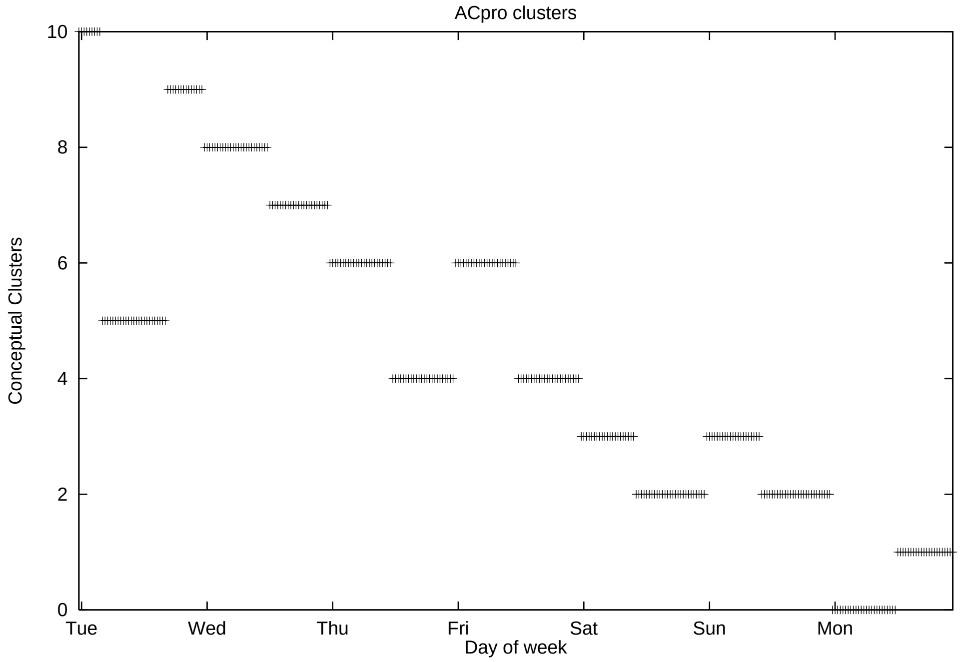

Figure 5: ACPro clustering on Elec2-1, run 3. Distribution of clusters over any given week.

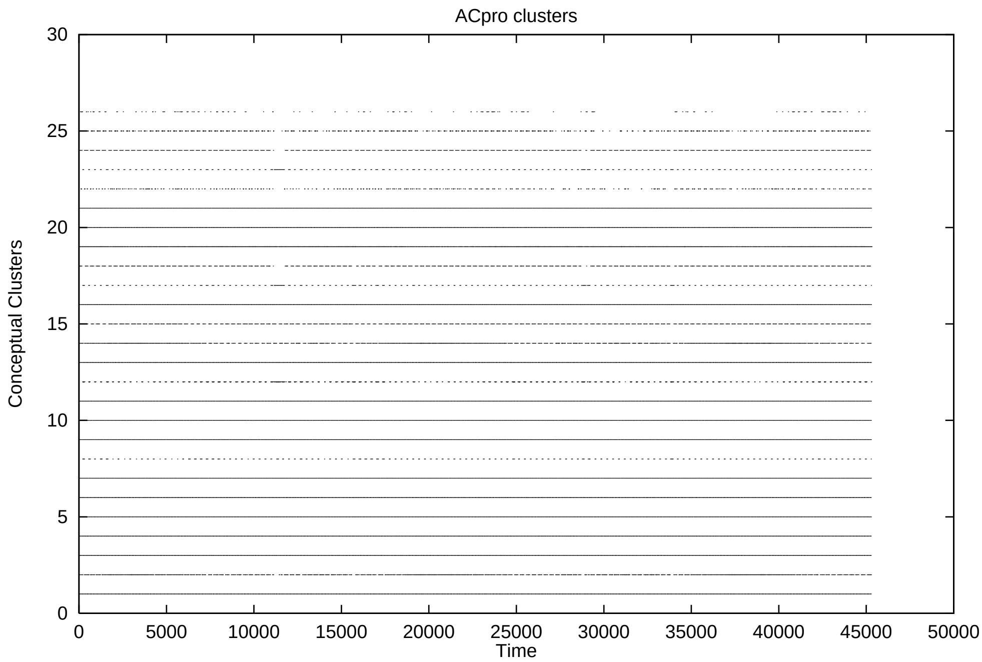

Figure 6: ACPro clustering on Elec2-2.

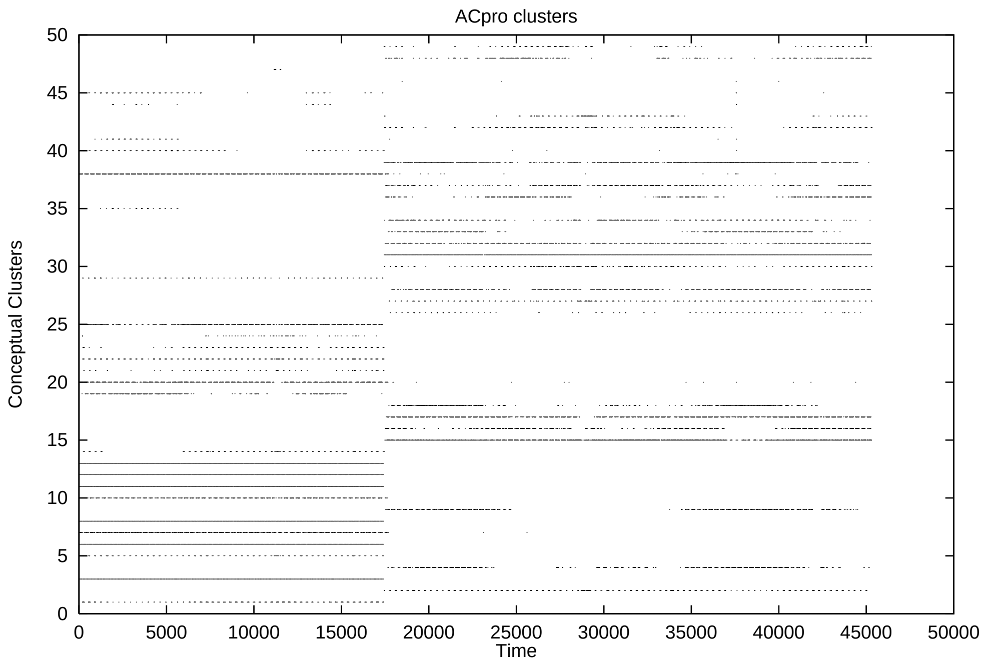

Figure 7: ACPro clustering on Elec2-3.

4 Discussion

The first set of experiments, on Elec1, were designed to investigate whether C4.5 could identify and exploit the hidden contexts identified by Splice-2. We found that adding the dataset order as an attribute substantially improved the C4.5 accuracy, but not to the level achieved by Splice-2. Also of interest, adding another *environmental attribute*[11], day length, did not provide much advantage to C4.5. In conclusion, C4.5 was unable to exploit the regularities in Elec1 exploited by Splice-2.

Correspondingly for Splice-2, the addition of an environmental attribute (data set order, day length) as a normal attribute, reduced the contextual clustering stability and accuracy achieved by Splice-2. This is not surprising as stable concepts incorporating these attributes will not generalise to other periods. Conversely, allowing Splice-2 to use either time or day length as a meta–attribute (i.e. as the data set order) led to large improvements in accuracy and clustering regularity. Finally, removing all order from the dataset destroyed Splice-2's advantage over C4.5.

The second set of experiments, on Elec2, investigated the impact of adding highly relevant attributes (for electricity demand). On the initial minimal dataset (Elec2-1) Splice-2 (with three contextual clusters) was

[11] An attribute in which hidden context is reflected.

able to achieve a 2.9% improvement in accuracy over C4.5 . Adding the additional attributes (Elec2-3) improved the C4.5 result by 2.0%; not enough to beat the initial SPLICE-2 result. When SPLICE-2 was run on the augmented dataset, however, its accuracy was reduced to only 0.4% better than the final C4.5 result. This level of improvement could be merely due to randomly partitioning a changing dataset.

These results demonstrate that SPLICE-2 was able to fill in for a missing attribute by utilising the dataset order. When the demand attributes were added, the increase in accuracy due to SPLICE-2 was reduced. In fact, adding the attributes reduced the absolute SPLICE-2 accuracy, suggesting that although the SPLICE-2 hidden contexts were more predictive than the new attributes, the new attributes confounded the SPLICE-2 contextual clustering process.

This damage to the contextual clustering process should be investigated in future work. Can it be alleviated by heavy pruning? By reducing the window size? By changes to the clustering mechanism? Or is it due due to overly strong attributes skewing the learning process.

The contextual clusters found by SPLICE-2 were very consistent for both Elec1-1 and Elec2-1 (the minimal datasets). These clusters were also apparently related to season; this makes sense according to domain experts. The addition of more attributes made the clusters less regular for repeated random runs of SPLICE-2 on the same data, however there were still frequent similarities in the clusters found in different runs. When SPLICE-2 was run with three contextual clusters it identified a possible shoulder season cluster, in which the first winter is similar to the beginning of the second and third winters, suggesting some evolution of the domain over time. Contextual clusters provide domain insight to the user, and hide classification details in the stable concepts.

Contrast the types of contextual clusters found by SPLICE-2 with the conceptual clusters found by ACPro. On the minimal datasets (Elec1-1 and Elec2-1) ACPro found very little structure. This is not surprising, as the two attributes are regularly distributed and thus can provide little insight from compression. When run, on the other hand, with the demand attributes, ACPro returns 50 clusters. These provides some insight to the domain, and make a very distinct break at the point at which the scheduled electricity transfer attribute becomes non-zero. Much of the structure in the ACPro clusters can be expected to be captured in a C4.5 model. Nonetheless, it is notable that the interesting ACPro structure only arises in the datasets where the SPLICE-2 contextual clustering does not converge well.

If we compare methods in terms of domain insight; C4.5 tends to have the smallest description size; SPLICE-2 has a combined tree size of no more than 2.4 times the C4.5 tree size and provides additional information about hidden context; C4.5 on-line has an enormous total description size (with 133 weeks worth of short term concepts) and provides very little insight.

Hence, for domain insight, the user should consider the use of C4.5 alone or in combination with SPLICE-2.

It is important to note that these results are inescapably qualified by the small number of independent attributes in the dataset. In future work it would be beneficial to look at a data type with a controllable quantity of attributes.

5 Conclusion

SPLICE-2 was compared, on a real world domain, with C4.5, an on-line learner, and an unsupervised learner. A broad range of Splice-2 characteristics were identified. These included a demonstration that neither C4.5 nor the unsupervised learner could be used to identify the hidden context. SPLICE-2 achieved higher accuracy than C4.5 except in the cases where the dataset order was destroyed, or where attributes were added that were too similar to *time*. Finally, the hidden contexts identified by SPLICE-2 were shown to be sensitive to the number of attributes in this domain.

6 Acknowledgements

Kim Horn, then of RMB Australia, suggested that I apply SPLICE to the electricity pricing domain. Kim also provided the the initial Elec-1 dataset.

Donald Michie provided much appreciated encouragement and enthusiasm. He made suggestions about both experimental design and report structure. He also read and commented on various drafts of the report.

My thesis supervisor, Claude Sammut, read and commented on various drafts of this report. As usual, the paper was much improved by his suggestions.

This work was supported by the Australian Research Council.

References

[1] Peter Cheesman, James Kelly, Matthew Self, John Stutz, Will Taylor, and Don Freeman. Autoclass: a Bayesian classification system. In *Proceedings of the Fifth International Conference on Machine Learning*, 1988.

[2] Michael B. Harries, Claude Sammut, and Kim Horn. Extracting hidden context. *Machine Learning*, 32:101–126, 1998.

[3] Miroslav Kubat, Robert C. Holte, and Stan Matwin. Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, 30:195–215, 1998.

[4] J. R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers Inc., San Mateo, California, 1993.

A Electricity Detailed Results

Following are logs of evaluation runs for C4.5, C4.5 on-line and SPLICE-2 on the electricity datasets.

A.1 Elec1-1

Dataset	elec1-1
C4.5 arguments	-s
C4.5 walkforward window sizes	336 672 1008 1344 1680 2016 2352 2688 3024 3360 100000
C4.5 walkforward step size	336
SPLICE-2 arguments	-c4.5arg '-s' -w 2000 -l 10 -r 2
Temporal cross validation settings	-c 336 -k 10

Data

Classes		Default
UP	DOWN	Accuracy
12507	16965	57.6%

C4.5

Xval Accuracies (%)										Average	Std Err
65.7	65.6	65.7	65.8	65.8	65.8	66	65.4	65.5	65.7	65.7%	0.1

Tree size (Nodes)
41

C4.5 on-line

Window Size	336	672	1008	1344	1680	2016	2352	2688	3024	3360	100000
Accuracy %	65.4	66.7	67	67.2	68.4	67.5	67.2	67.5	67.5	67.4	65.3

SPLICE-2

Xval Accuracies (%)										Average	Std Err
68.4	68.8	69.5	69	69.3	69.5	69.2	69.7	68.6	68.9	69.1%	0.1

Figure 8 shows SPLICE-2 clusters from 10 random initializations, given ten iterations of contextual clustering.

Node count by Context		Total
73	42	115
22	46	68
55	39	94
54	46	100
64	43	107
52	41	93
46	41	87
49	42	91
68	44	112
64	42	106
Average		97.3

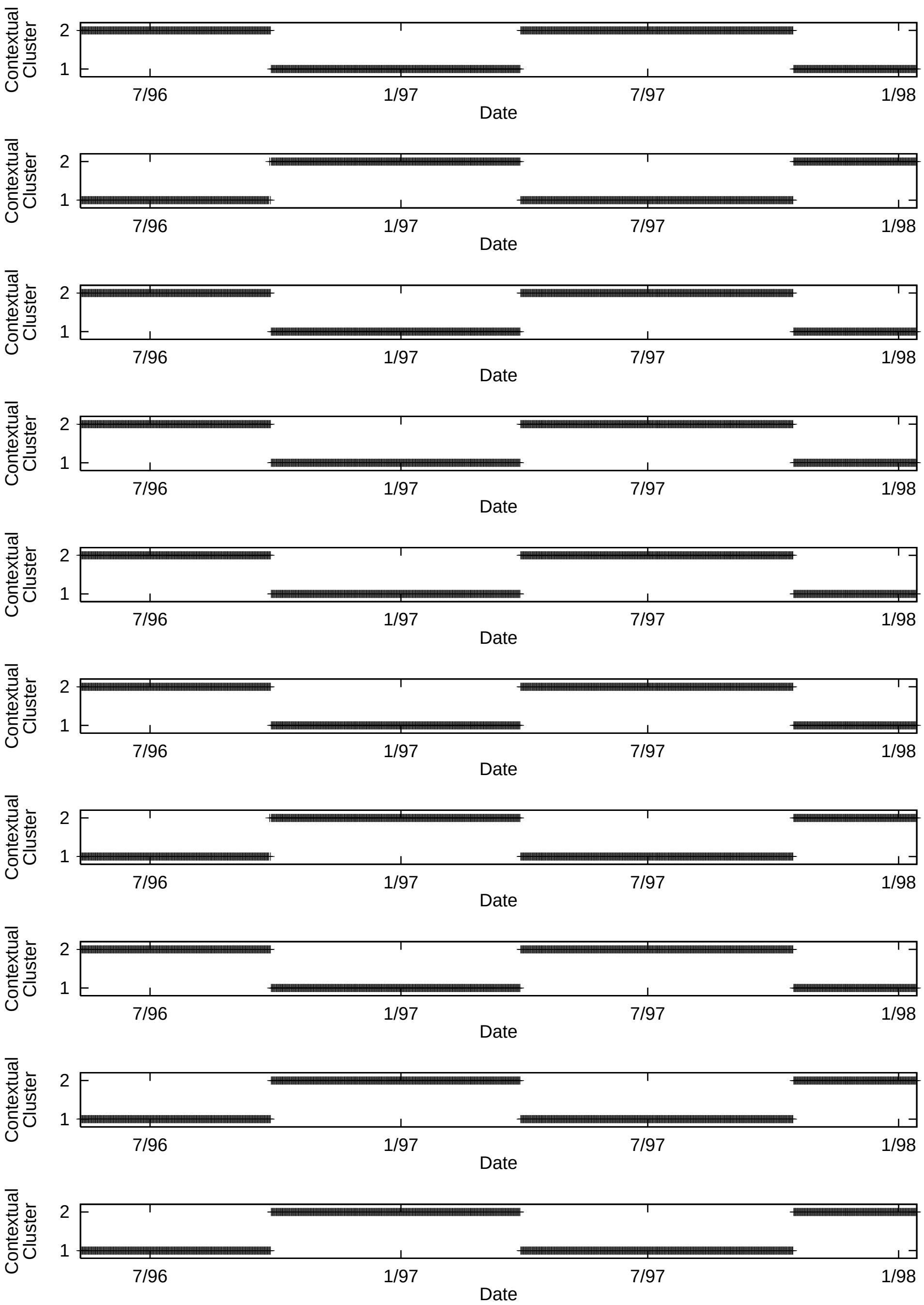

Figure 8: SPLICE-2 clustering on Elec1-1, given ten iterations of contextual clustering.

SPLICE-2 Iterations

Iterations	Xval Accuracies (%)										Average	Std Err
0	66.7	66.0	65.9	66.3	66.0	65.6	65.9	66.4	66	66.2	66.1%	0.1
1	66.3	66.8	67.5	67.5	66.8	67.1	67.4	67.6	67.2	67.3	67.2%	0.1
2	67.8	68.5	68.9	68.2	67.8	68.0	68.3	68.7	67.7	68.4	68.2%	0.1
3	68.1	68.9	69.4	68.5	67.8	69.1	68.5	68.5	68.2	68.9	68.6%	0.1
4	68.0	68.8	69.5	68.7	68.7	69.6	69.0	68.9	68.1	68.8	68.8%	0.2
5	68.1	68.7	69.5	68.8	69.2	69.5	69.0	69.4	68.5	68.8	69.0%	0.1
6	68.1	68.7	69.4	68.9	69.3	69.6	68.9	69.7	68.7	68.9	69.0%	0.1
7	68.0	68.6	69.3	68.9	69.3	69.6	69.0	69.7	68.7	68.9	69.0%	0.15
8	68.3	68.8	69.5	69.0	69.2	69.6	69.0	69.6	68.6	68.8	69.0%	0.13
9	68.5	68.7	69.5	68.9	69.2	69.5	69.1	69.7	68.5	68.7	69.0%	0.13
10	68.4	68.8	69.5	69	69.3	69.5	69.2	69.7	68.6	68.9	69.1%	0.1

Figure 9 shows two clusters from 10 random dataset partitionings. Figures 10, 11, 12, 13, 14, 15, 16, 17, 18, and 8 show the impact of iterative the SPLICE-2 clustering algorithm first only one time, then twice, etc.

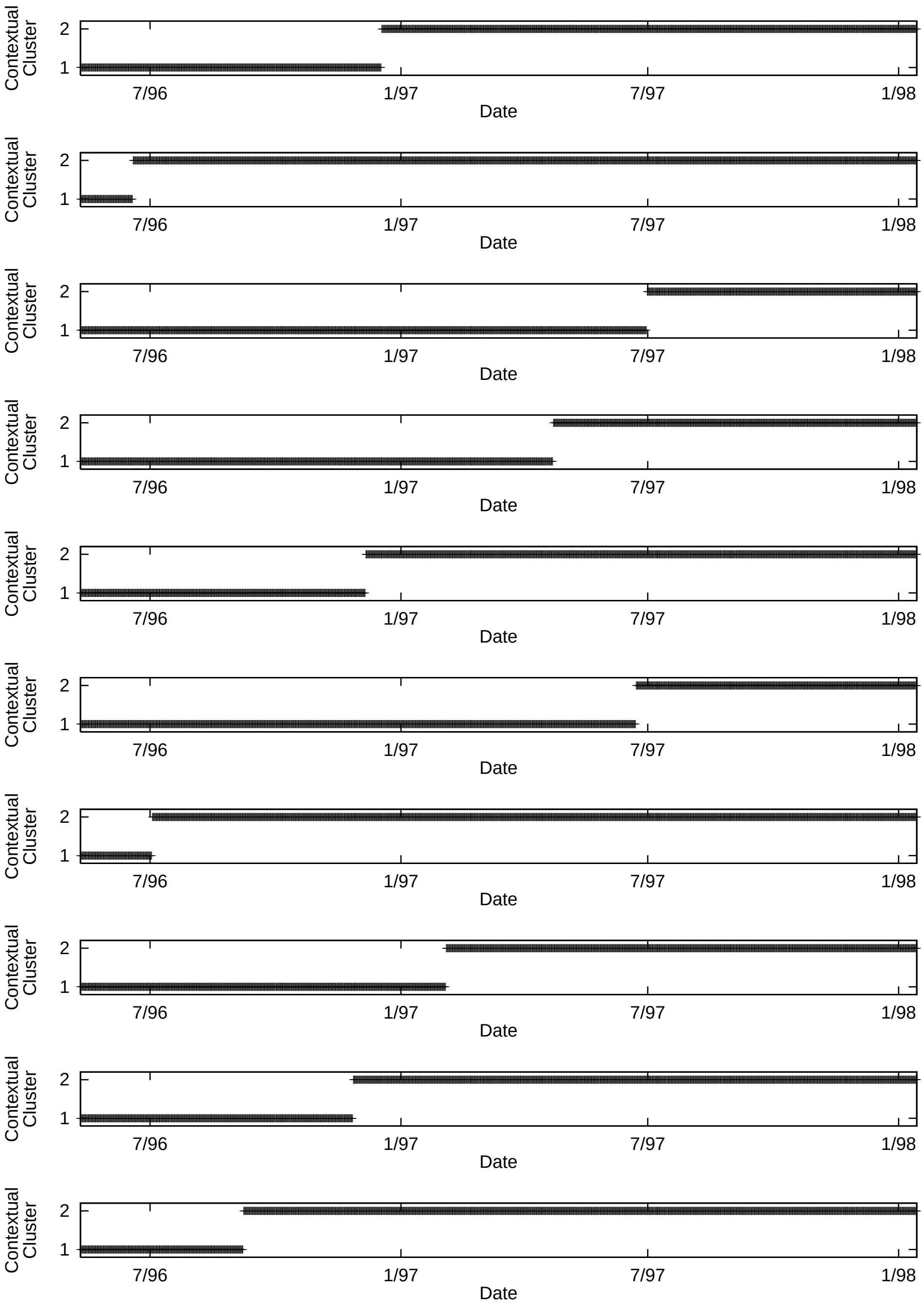

Figure 9: SPLICE-2 clustering on Elec1-1, with no clustering iterations.

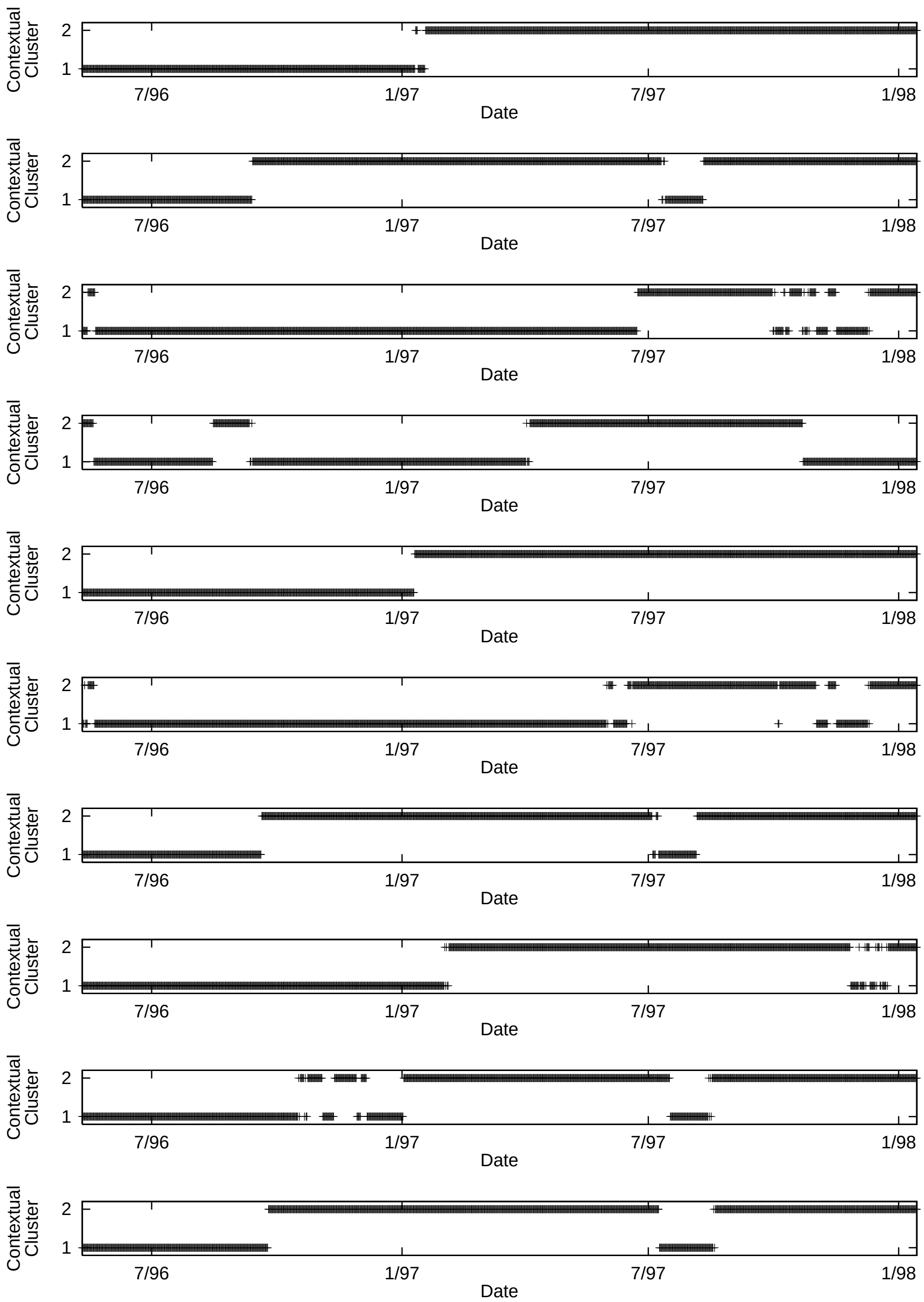

Figure 10: Splice-2 clustering on Elec1-1, given 1 iteration of contextual clustering.

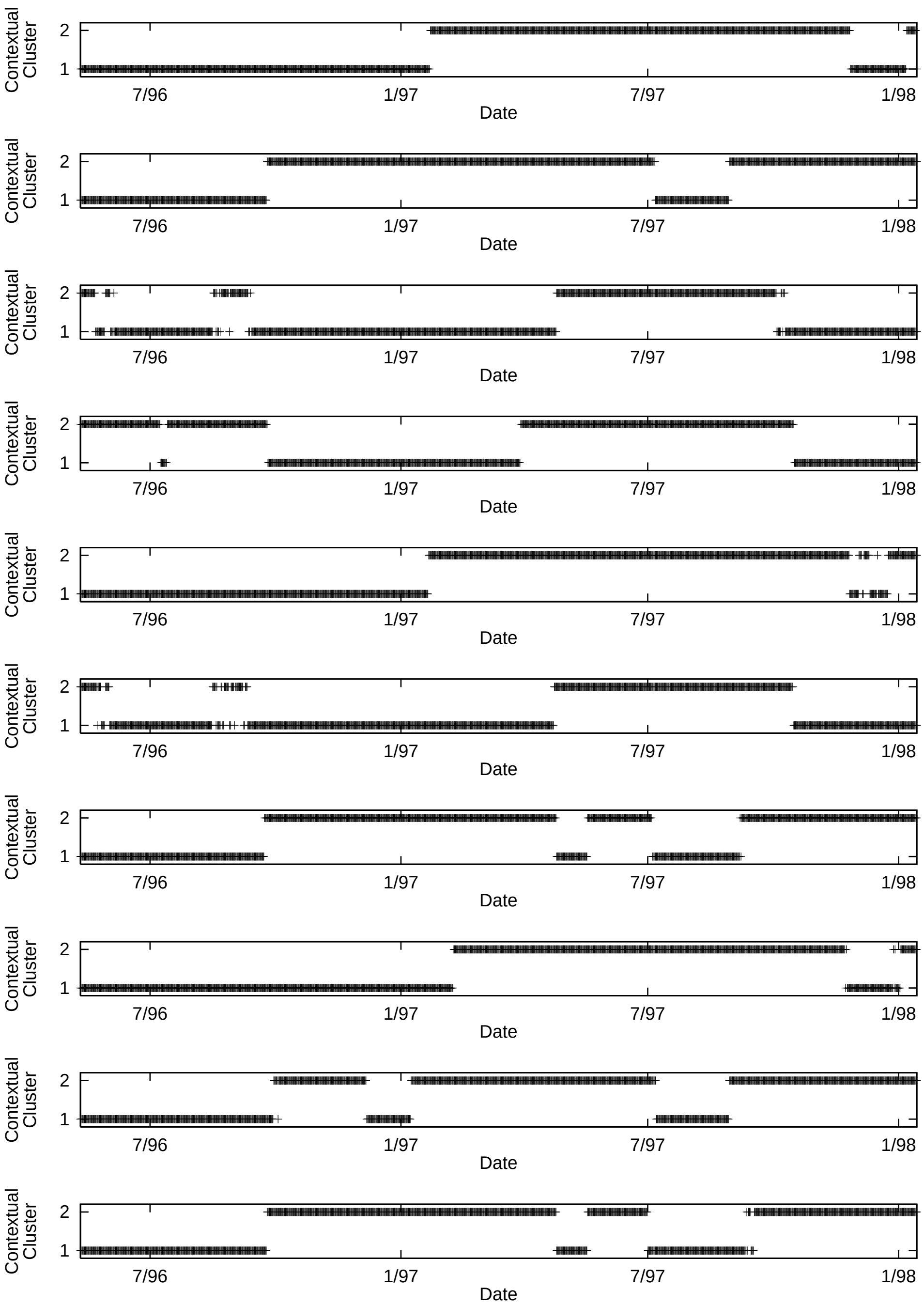

Figure 11: SPLICE-2 clustering on Elec1-1, given two iterations of contextual clustering

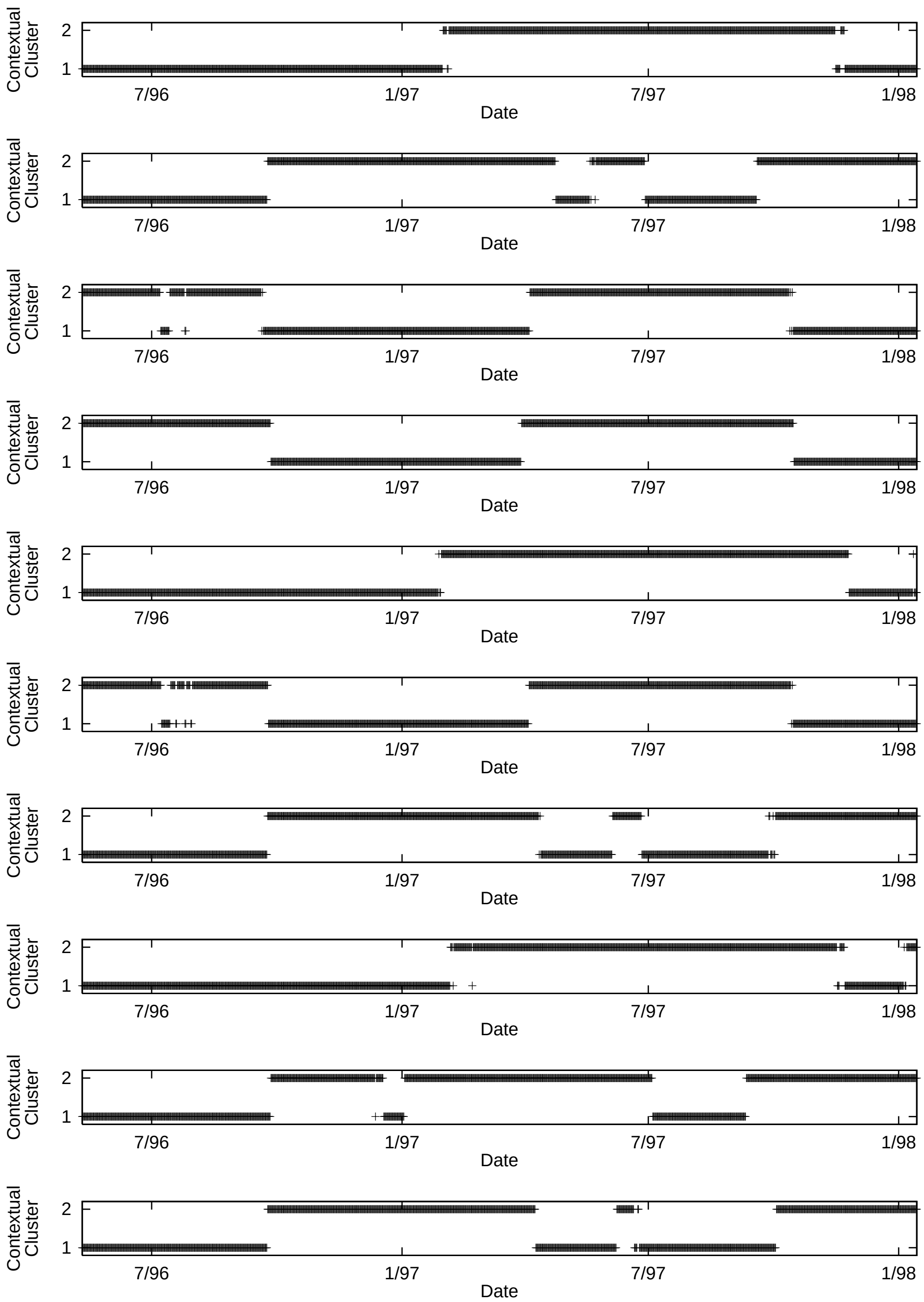

Figure 12: SPLICE-2 clustering on Elec1-1, given three iterations of contextual clustering.

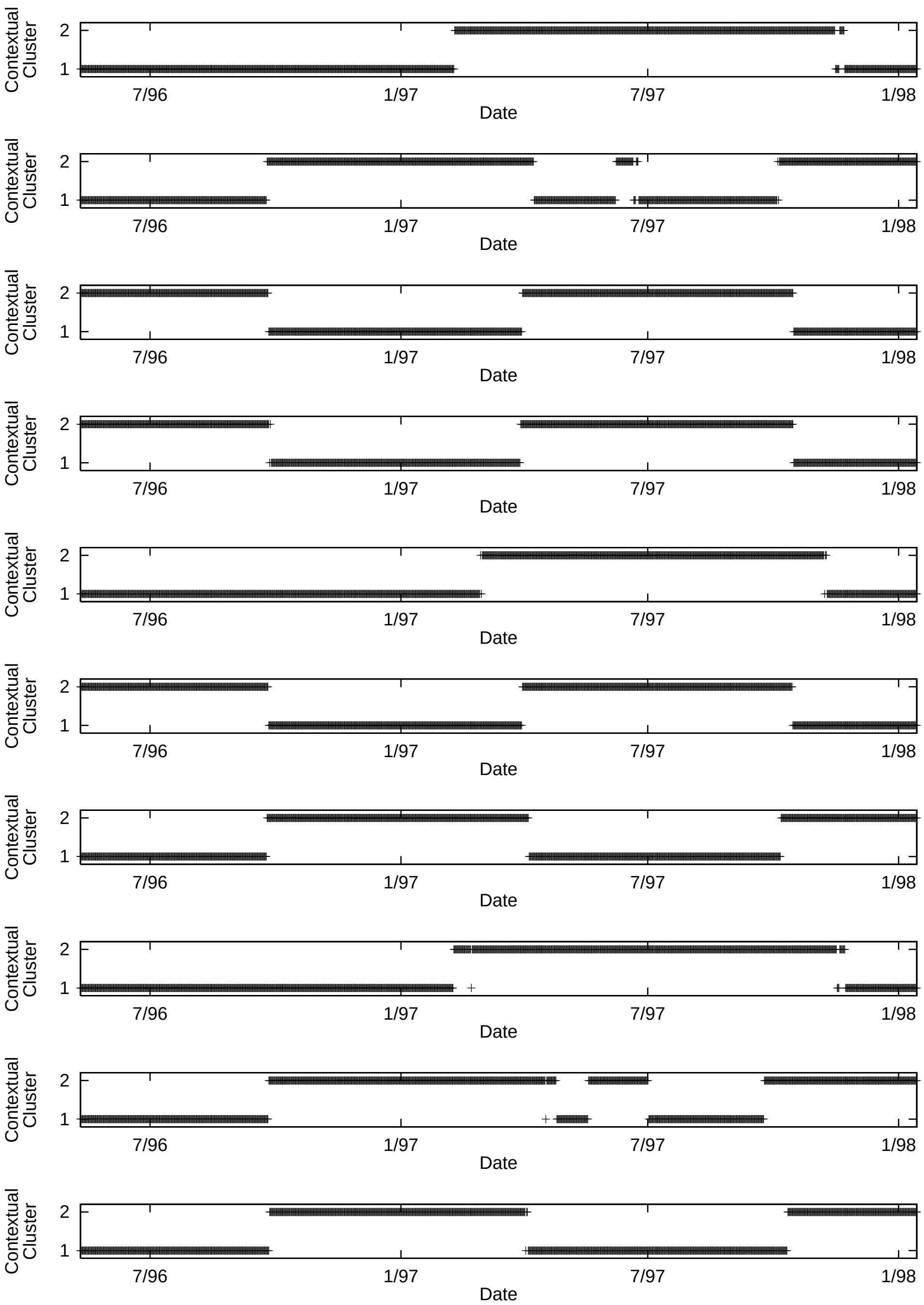

Figure 13: SPLICE-2 clustering on Elec1-1, given four iterations of contextual clustering.

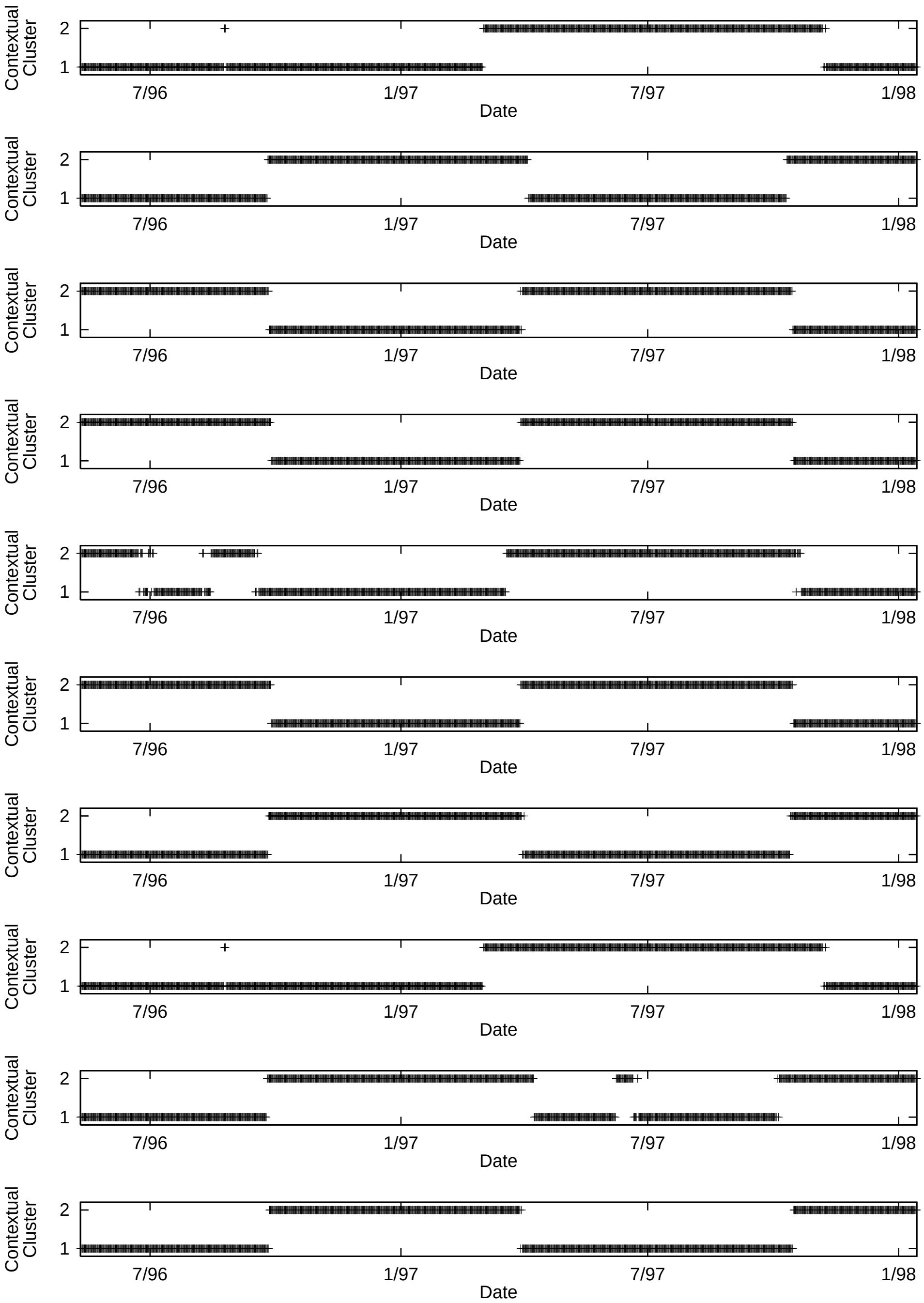

Figure 14: SPLICE-2 clustering on Elec1-1, given five iterations of contextual clustering.

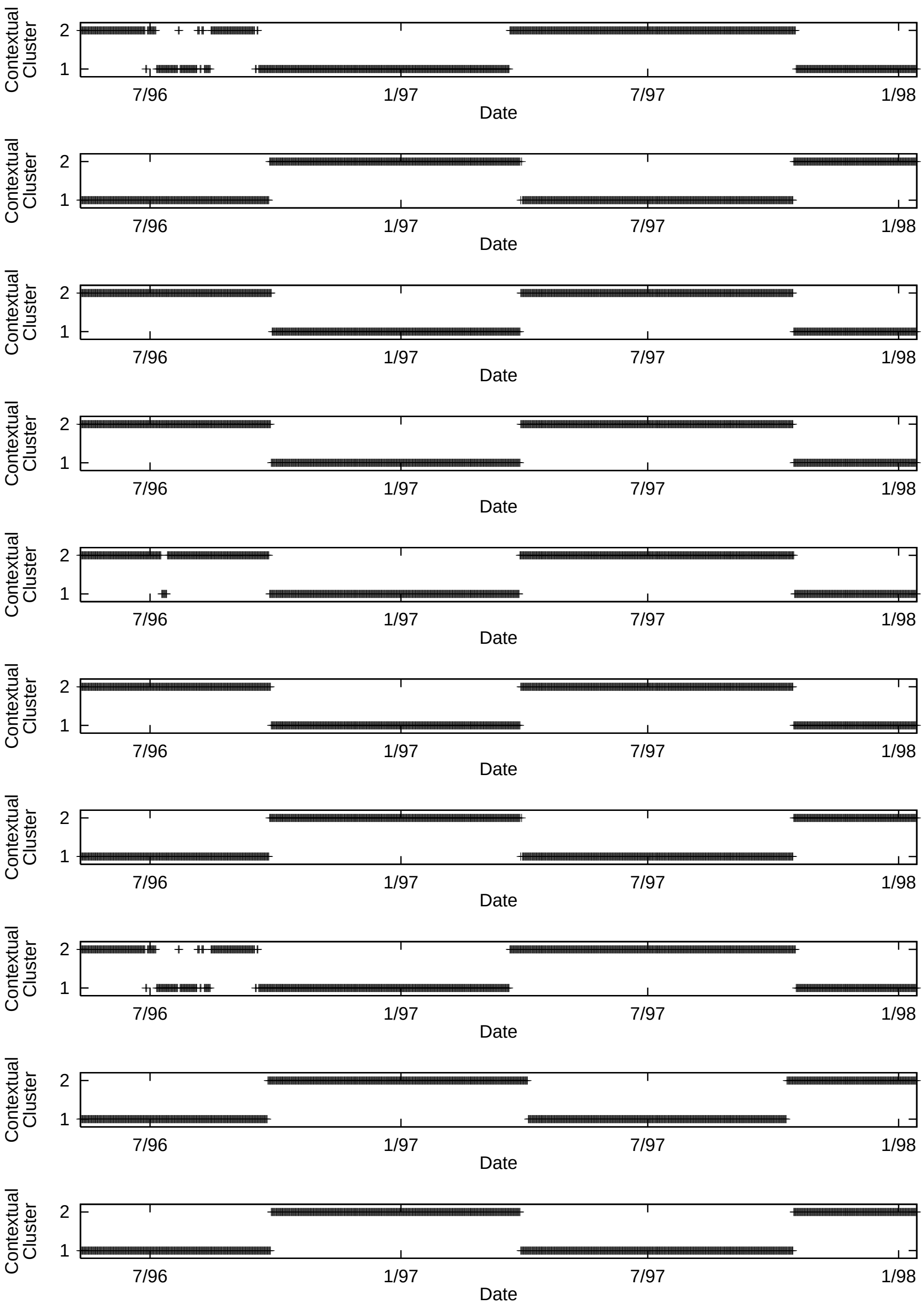

Figure 15: SPLICE-2 clustering on Elec1-1, given six iterations of contextual clustering.

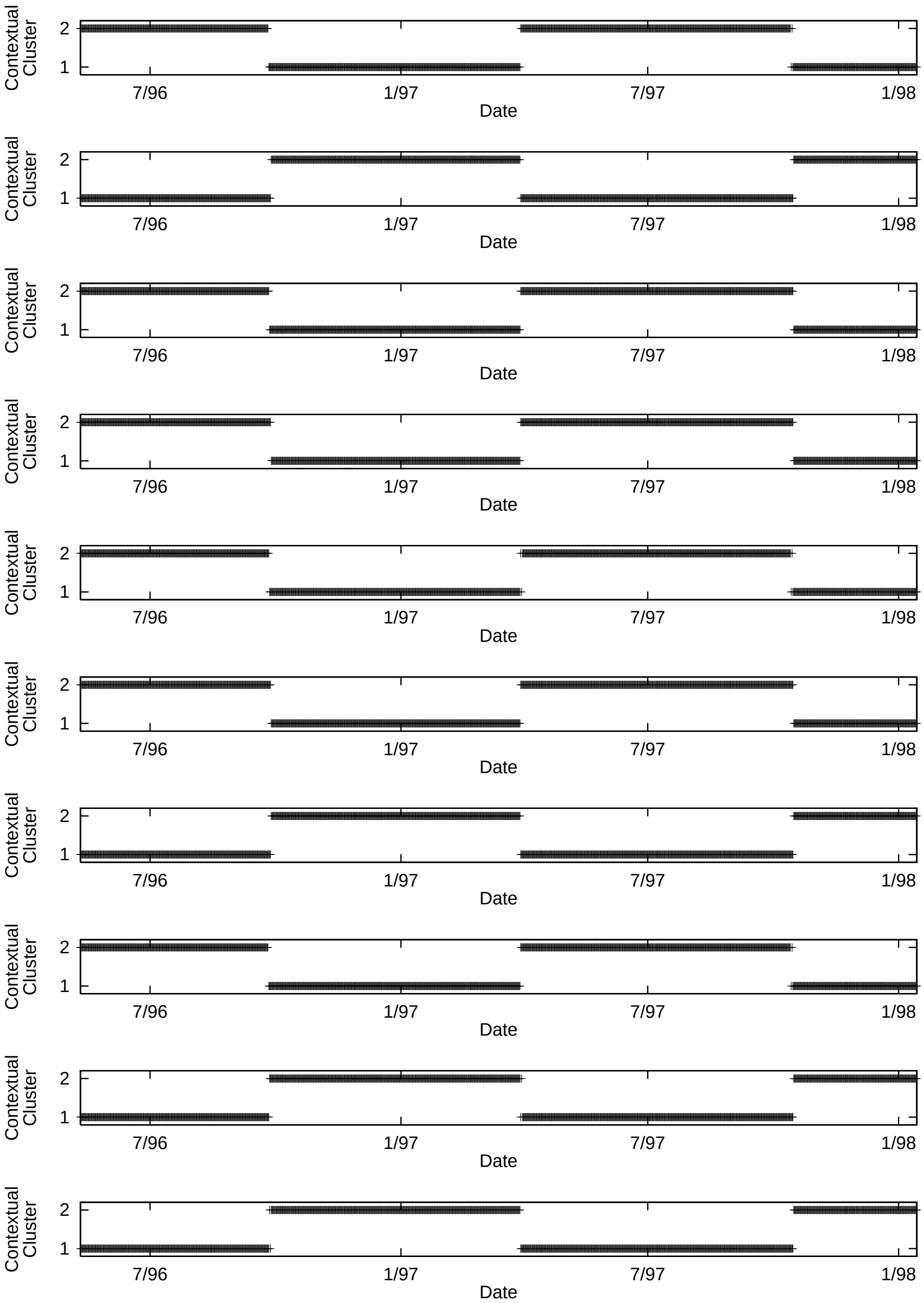

Figure 16: SPLICE-2 clustering on Elec1-1, given seven iterations of contextual clustering.

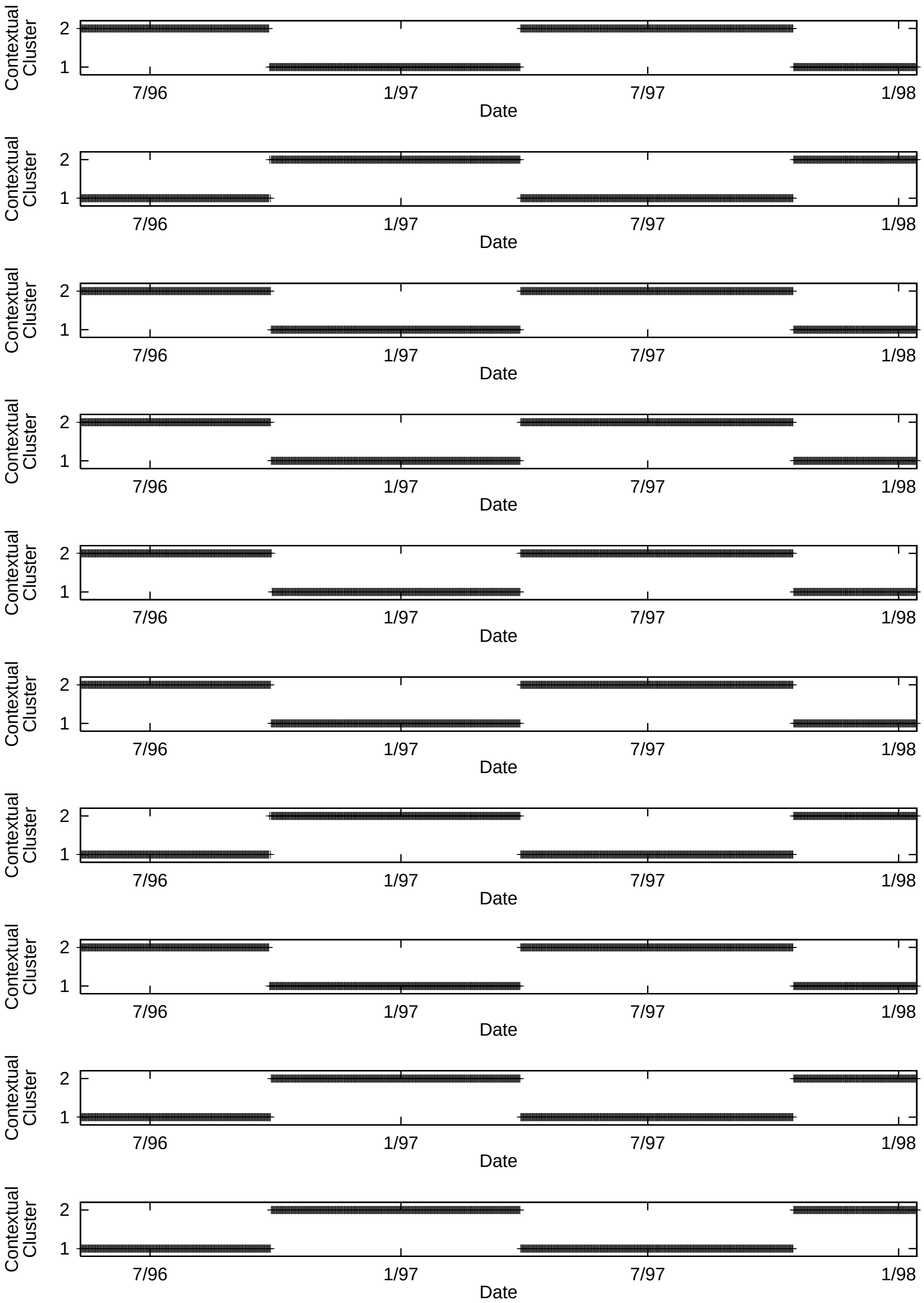

Figure 17: SPLICE-2 clustering on Elec1-1, given eight iterations of contextual clustering.

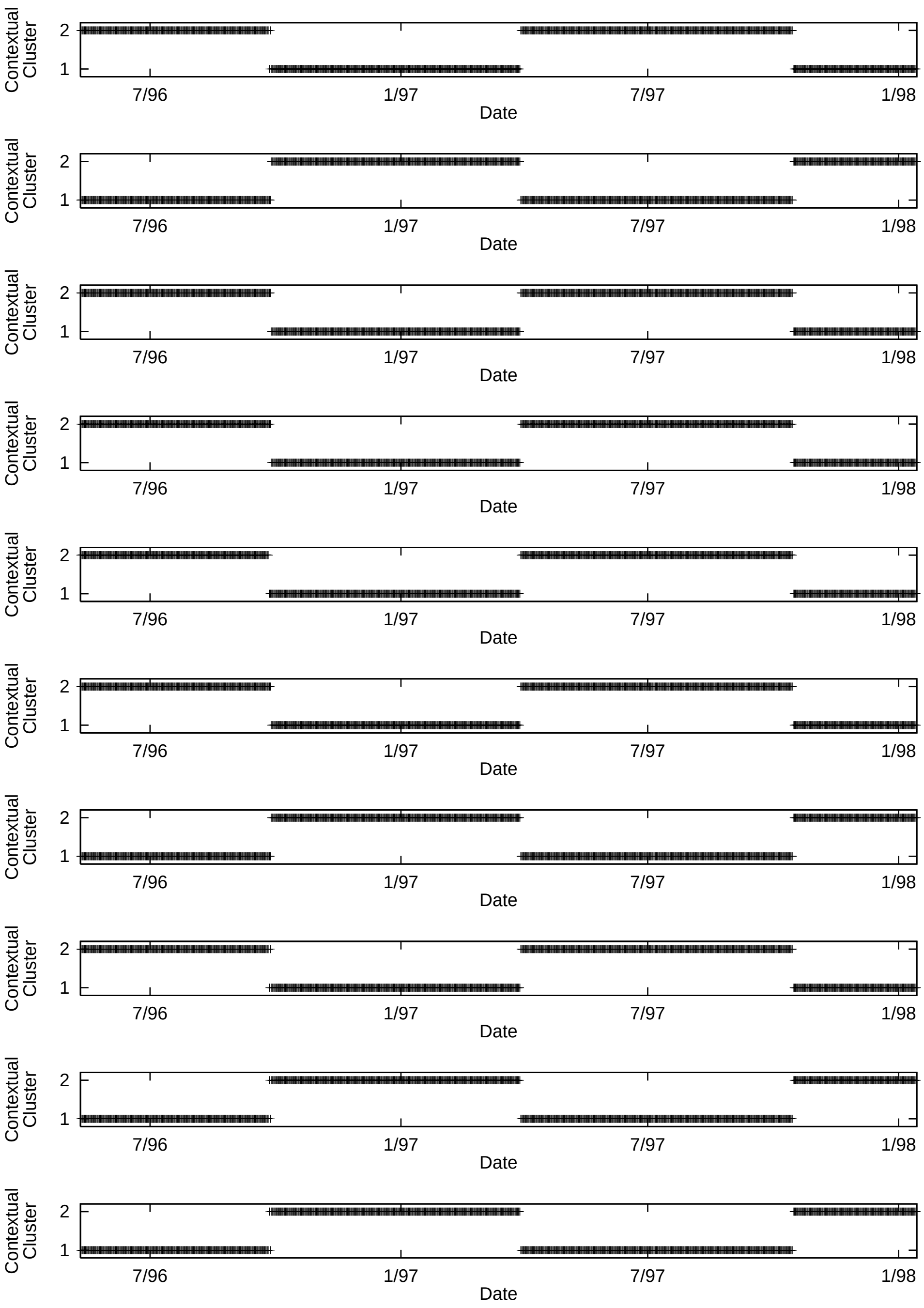

Figure 18: SPLICE-2 clustering on Elec1-1, given nine iterations of contextual clustering.

A.2 Elec1-2

Dataset	elec1-2
C4.5 arguments	-s
C4.5 walkforward window sizes	336 672 1008 1344 1680 2016 2352 2688 3024 3360 100000
C4.5 walkforward step size	336
SPLICE-2 arguments	-c4.5arg '-s' -w 2000 -l 10 -r 2
Temporal cross validation settings	-c 336 -k 10

Data

Classes		Default
UP	DOWN	Accuracy
12507	16965	57.6%

C4.5

Xval Accuracies (%)										Average	Std Err
68	67.7	67.8	67.7	68.6	67.9	68.1	68.1	68.5	68.6	68.1%	0.11

Tree size (Nodes)
1535

C4.5 on-line

Window Size	336	672	1008	1344	1680	2016	2352	2688	3024	3360	100000
Accuracy %	61.5	61.6	62.9	64.3	63.4	64.3	64	64.5	64.2	64.4	65.2

SPLICE-2

Xval Accuracies (%)										Average	Std Err
68	67.5	67.9	67.6	67.7	67	68.2	67.8	67.4	67.5	67.7%	0.10

Figure 19 shows SPLICE-2 clusters from 10 random initializations.

Node count by Context		Total
850	1185	2035
287	1533	1820
1206	696	1902
1077	840	1917
715	1251	1966
1183	786	1969
283	1498	1781
964	936	1900
715	1251	1966
516	1470	1986
Average		1924.2

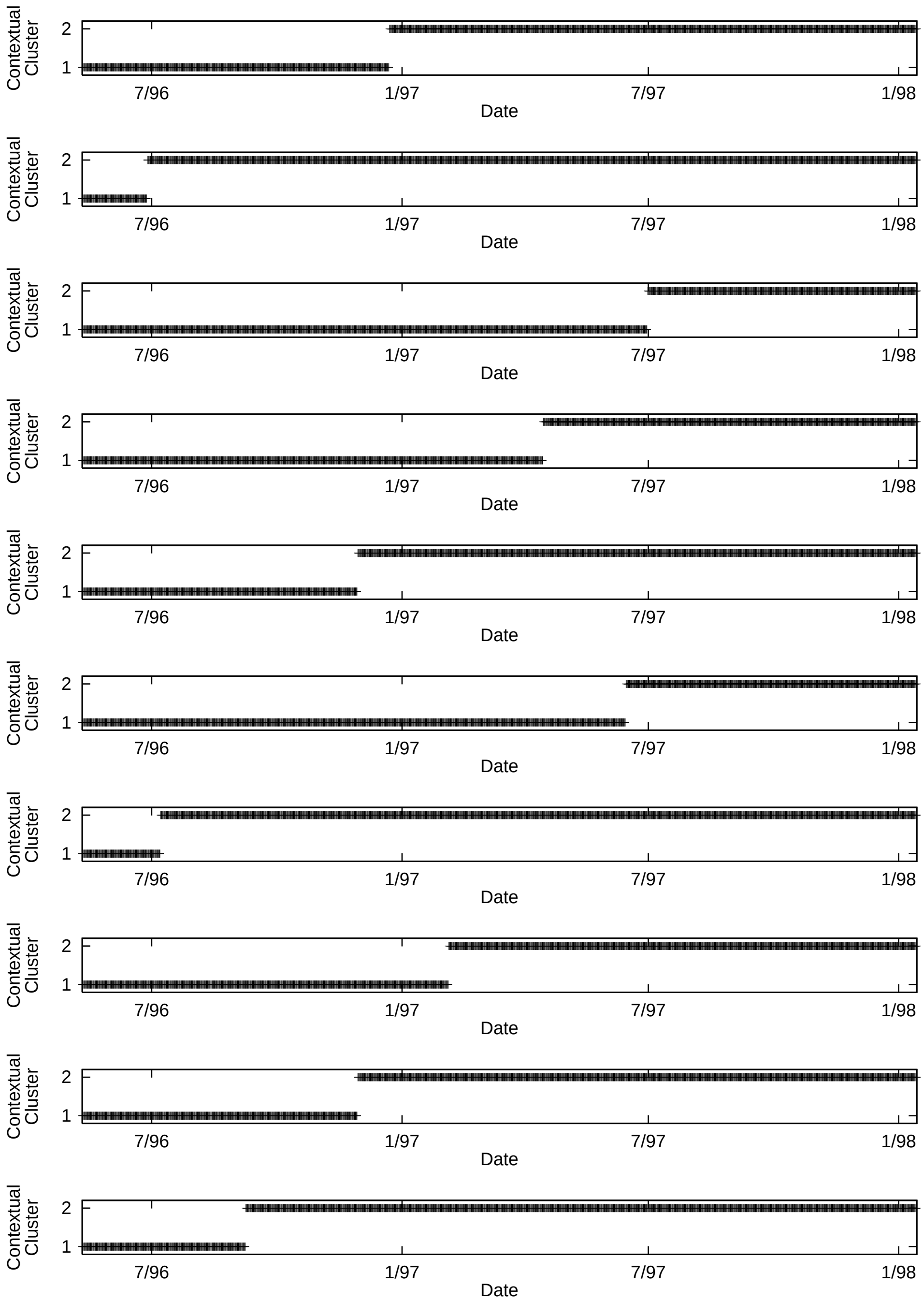

Figure 19: SPLICE-2 clustering on Elec1-2.

A.3 Elec1-3

Dataset	elec1-3
C4.5 arguments	-s
C4.5 walkforward window sizes	336 672 1008 1344 1680 2016 2352 2688 3024 3360 100000
C4.5 walkforward step size	336
SPLICE-2 arguments	-c4.5arg '-s' -w 2000 -l 10 -r 2
Temporal cross validation settings	-c 336 -k 10

Data

Classes		Default
UP	DOWN	Accuracy
12507	16965	57.6%

C4.5

Xval Accuracies (%)										Average	Std Err
66.1	65.3	66.3	66.5	66	66.9	66.2	66.1	65.8	65.6	66.1%	0.14

Tree size (Nodes)
1045

C4.5 on-line

Window Size	336	672	1008	1344	1680	2016	2352	2688	3024	3360	100000
Accuracy %	65.2	64.8	64.9	65.4	64.9	64.9	64.9	65.3	65.2	65.3	63.6

SPLICE-2

Xval Accuracies (%)										Average	Std Err
63.6	65.3	64.7	64.5	64.8	65.6	65.5	66.2	64.7	65.3	65.0%	0.22

Figure 20 shows SPLICE-2 clusters from 10 random initializations.

Node count by Context		Total
960	830	1790
298	1095	1393
1203	885	2088
1021	823	1844
964	844	1808
956	881	1837
301	1106	1407
1103	950	2053
959	841	1800
645	1007	1652
Average		1767.2

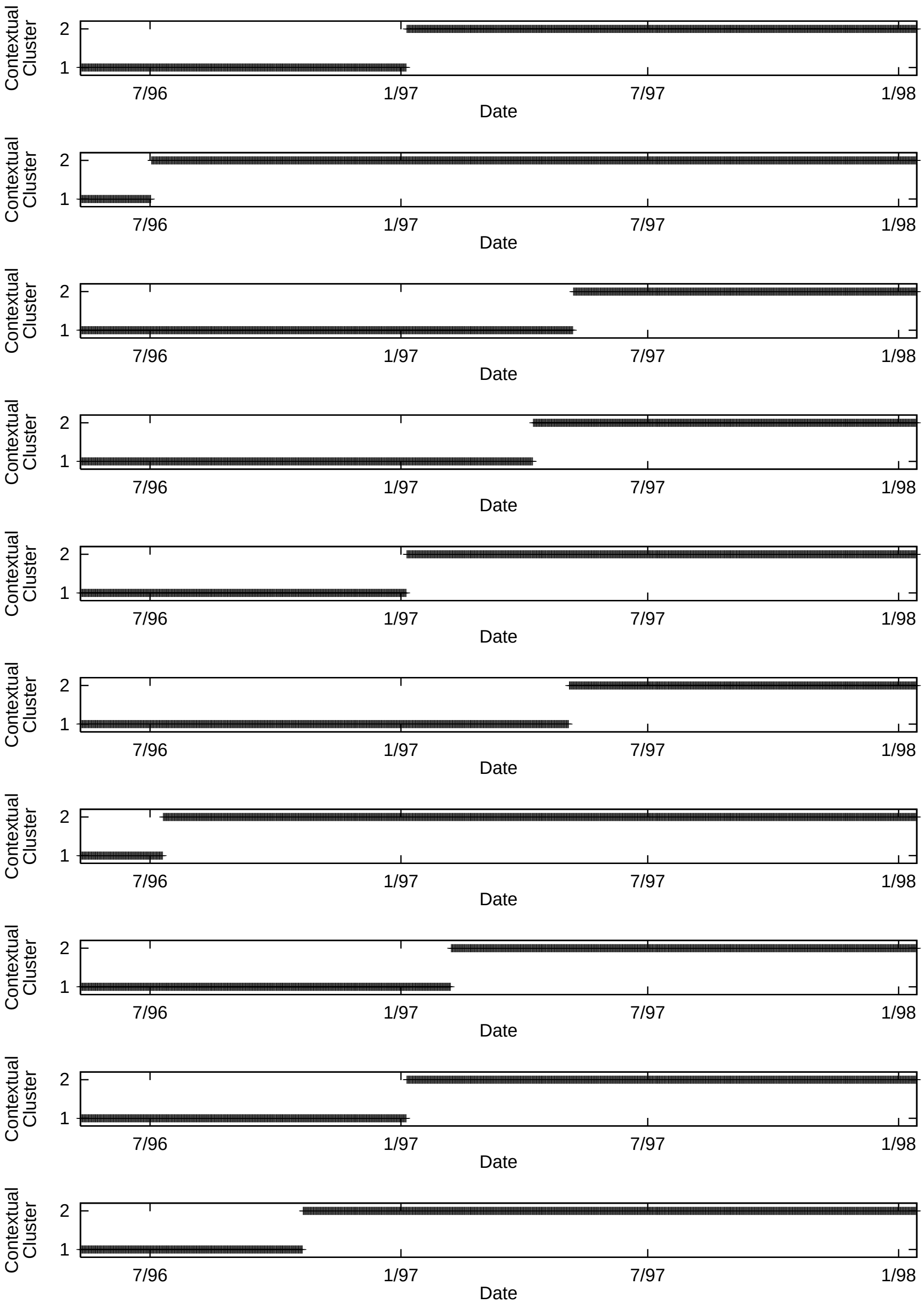

Figure 20: SPLICE-2 clustering on Elec1-3.

A.4 Elec1-4

Dataset	elec1-4
C4.5 arguments	-s
C4.5 walkforward window sizes	336 672 1008 1344 1680 2016 2352 2688 3024 3360 100000
C4.5 walkforward step size	336
Splice-2 arguments	-c4.5arg '-s' -w 2000 -l 10 -r 2
Temporal cross validation settings	-c 336 -k 10

Data

Classes		Default
UP	DOWN	Accuracy
12507	16965	57.6%

C4.5

Xval Accuracies (%)										Average	Std Err
65.8	65.6	65.5	65.7	65.7	65.5	65.8	65.8	65.4	65.8	65.7%	0.05

Tree size (Nodes)
41

C4.5 on-line

Window Size	336	672	1008	1344	1680	2016	2352	2688	3024	3360	100000
Accuracy %	61.7	62.5	65.2	65.9	67	66.8	67.8	68.1	68	68.3	65.3

Splice-2

Xval Accuracies (%)										Average	Std Err
69.5	69.4	69.6	69.6	69.2	69.4	69.2	69.2	69.5	69.6	69.4%	0.05

Figure 21 shows Splice-2 clusters from 10 random initializations.

Node count by Context		Total
52	33	85
52	33	85
52	33	85
52	33	85
52	33	85
52	33	85
52	33	85
52	33	85
52	33	85
52	33	85
Average		85.0

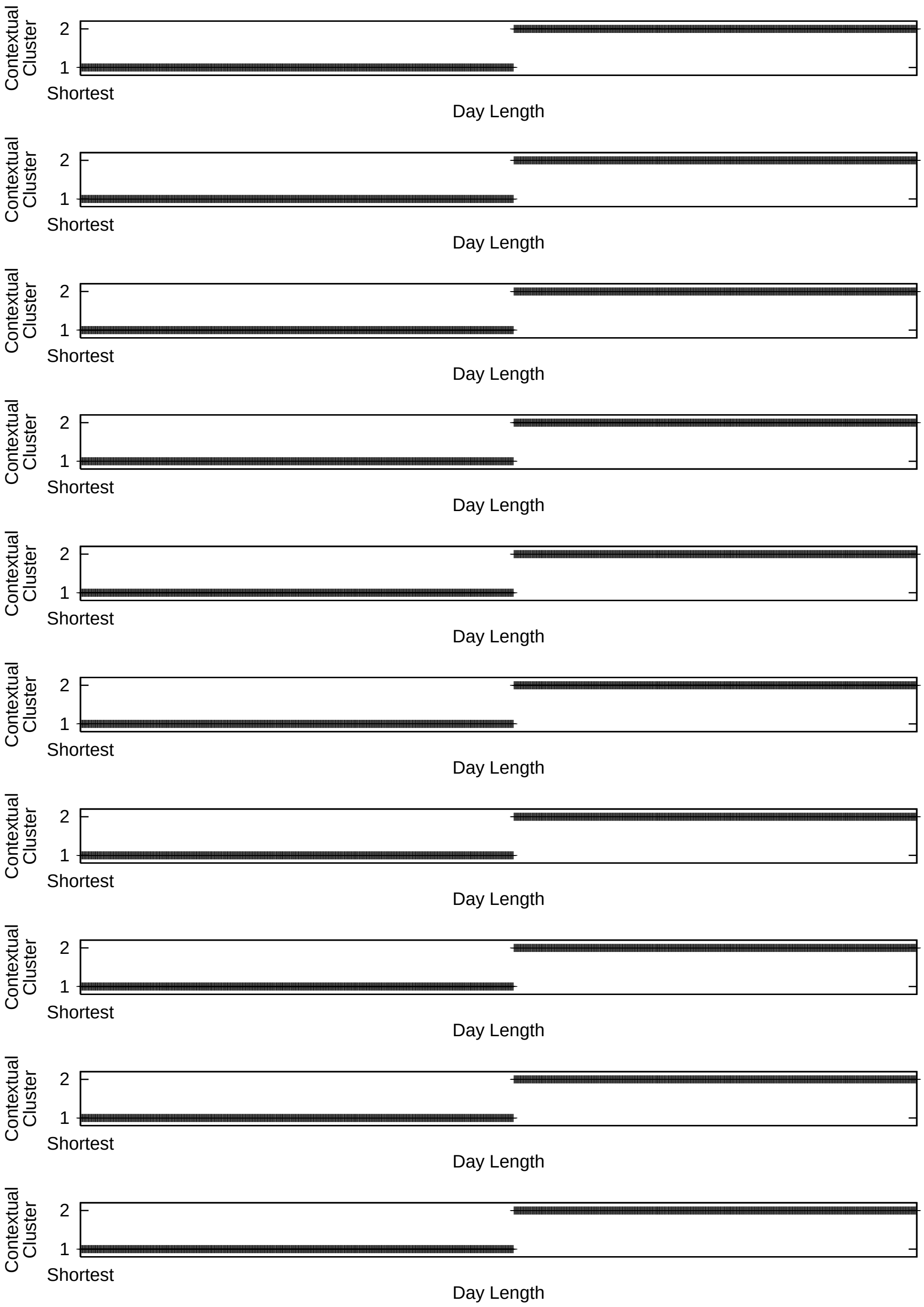

Figure 21: SPLICE-2 clustering on Elec1-4.

A.5 Elec1-5

Dataset	elec1-5
C4.5 arguments	-s
C4.5 walkforward window sizes	336 672 1008 1344 1680 2016 2352 2688 3024 3360 100000
C4.5 walkforward step size	336
SPLICE-2 arguments	-c4.5arg '-s' -w 2000 -l 10 -r 2
Temporal cross validation settings	-c 336 -k 10

Data

Classes		Default
UP	DOWN	Accuracy
12507	16965	57.6%

C4.5

Xval Accuracies (%)										Average	Std Err
66.2	66.4	66.3	66.1	66.2	66.2	66.3	66.3	66.5	66.1	66.3%	0.04

Tree size (Nodes)
41

C4.5 on-line

Window Size	336	672	1008	1344	1680	2016	2352	2688	3024	3360	100000
Accuracy %	62	62.7	63.2	63.8	64.3	64.8	64.8	65.2	65.2	65.4	65.9

SPLICE-2

Xval Accuracies (%)										Average	Std Err
65.9	66.3	65.9	65.9	65.9	66.1	66.2	66.1	66.2	66.1	66.1%	0.05

Figure 22 shows SPLICE-2 clusters from 10 random initializations.

Node count by Context		Total
54	40	94
41		41
47	39	86
38	45	83
32	36	68
41		41
52	33	85
50	53	103
36	46	82
43	48	91
Average		77.4

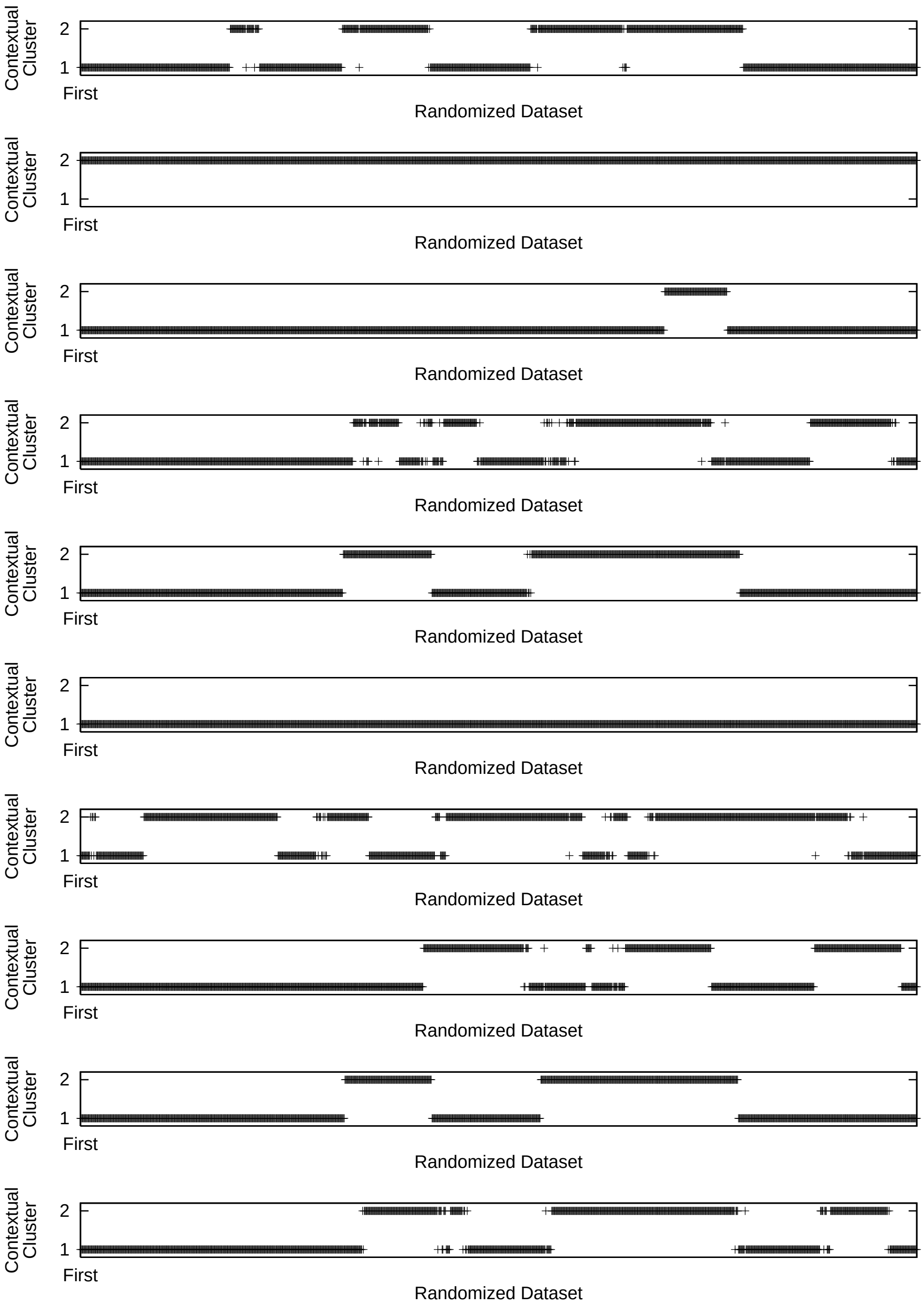

Figure 22: Splice-2 clustering on Elec1-5.

A.6 Elec2-1

Dataset	elec2-1
C4.5 arguments	-s
C4.5 walkforward window sizes	336 672 1008 1344 1680 2016 2352 2688 3024 3360 100000
C4.5 walkforward step size	336
Splice-2 arguments	-c4.5arg '-s' -w 2000 -l 10 -r 2
Temporal cross validation settings	-c 336 -k 10

Data

Classes		Default
UP	DOWN	Accuracy
19237	26075	57.5%

C4.5

Xval Accuracies (%)										Average	Std Err
65.7	65.2	65.8	65.7	65.9	65.2	65.2	65.7	65.2	65.7	65.5%	0.09

Tree size (Nodes)
54

C4.5 on-line

Window Size	336	672	1008	1344	1680	2016	2352	2688	3024	3360	100000
Accuracy %	65.7	66.5	67.7	67.5	68.4	67.5	67.3	67.5	67.6	67.3	65

Splice-2

Xval Accuracies (%)										Average	Std Err
67.6	66.8	67.9	67.5	67.7	66.8	66.8	66.9	66.8	68.1	67.3%	0.16

Figure 23 shows Splice-2 clusters from 10 random initializations.

Node count by Context		Total
24	66	90
45	41	86
24	66	90
49	34	83
24	61	85
24	66	90
45	41	86
24	66	90
24	66	90
54	41	95
Average		88.5

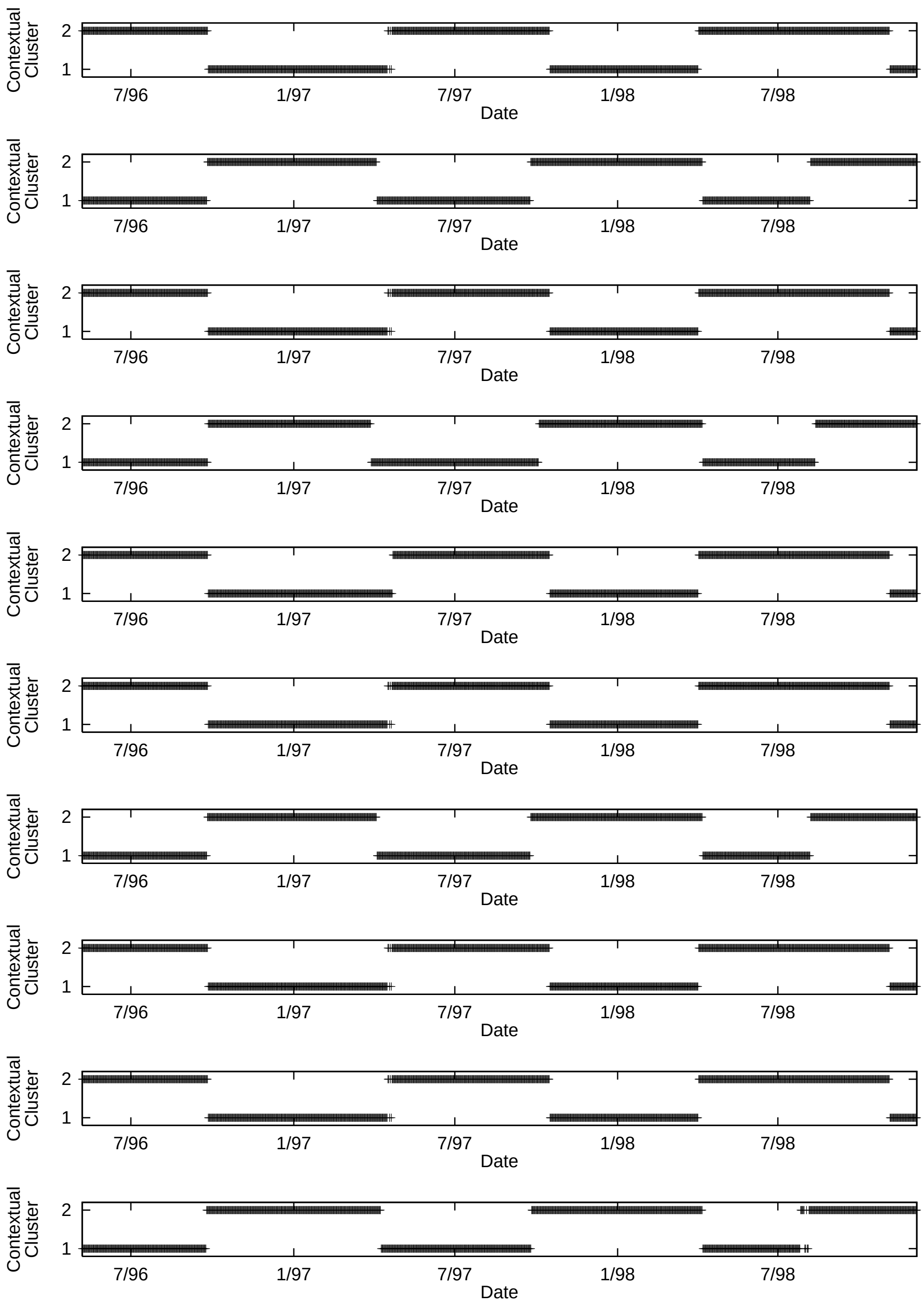

Figure 23: SPLICE-2 clustering on Elec2-1.

A.7 Elec2-2

Dataset	elec2-2
C4.5 arguments	-s
C4.5 walkforward window sizes	336 672 1008 1344 1680 2016 2352 2688 3024 3360 100000
C4.5 walkforward step size	336
SPLICE-2 arguments	-c4.5arg '-s' -w 2000 -l 10 -r 2
Temporal cross validation settings	-c 336 -k 10

Data

Classes		Default
UP	DOWN	Accuracy
19237	26075	57.5%

C4.5

Xval Accuracies (%)										Average	Std Err
67.5	66.9	67.1	67.5	67.6	66.9	66.9	67.7	66.9	67.5	67.2%	0.10

Tree size (Nodes)
386

C4.5 on-line

Window Size	336	672	1008	1344	1680	2016	2352	2688	3024	3360	100000
Accuracy %	66.5	67.2	68.1	68.2	68.1	67.8	68	68.1	68.1	68.4	66.1

SPLICE-2

Xval Accuracies (%)										Average	Std Err
68.1	66.6	68.1	68	68.4	66.6	66.6	68.5	66.6	68.3	67.6%	0.26

Figure 24 shows SPLICE-2 clusters from 10 random initializations.

Node count by Context		Total
292	228	520
205	359	564
258	217	475
286	245	531
292	228	520
282	239	521
187	389	576
259	274	533
266	290	556
228	337	565
Average		536.1

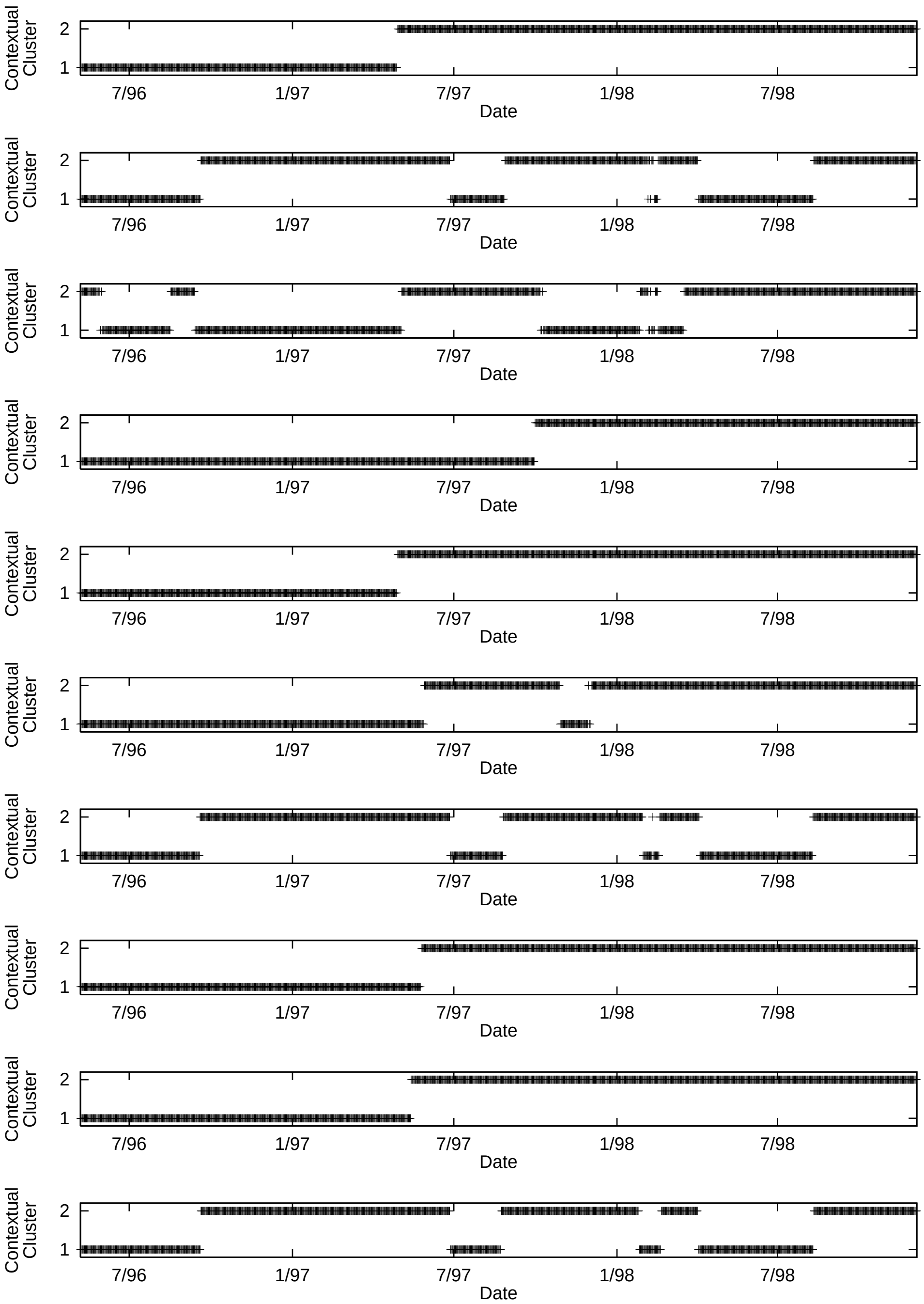

Figure 24: SPLICE-2 clustering on Elec2-2.

A.8 Elec2-3

Dataset	elec2-3
C4.5 arguments	-s
C4.5 walkforward window sizes	336 672 1008 1344 1680 2016 2352 2688 3024 3360 100000
C4.5 walkforward step size	336
SPLICE-2 arguments	-c4.5arg '-s' -w 2000 -l 10 -r 2
Temporal cross validation settings	-c 336 -k 10

Data

Classes		Default
UP	DOWN	Accuracy
19237	26075	57.5%

C4.5

Xval Accuracies (%)										Average	Std Err
68.1	66.9	68.1	67.9	67.8	66.9	66.9	67.9	66.9	68.1	67.5%	0.17

Tree size (Nodes)
1221

C4.5 on-line

Window Size	336	672	1008	1344	1680	2016	2352	2688	3024	3360	100000
Accuracy %	66	67.3	67.2	67.4	67.7	67.5	67.7	67.6	67.5	67.7	66.8

SPLICE-2

Xval Accuracies (%)										Average	Std Err
67.8	66.9	68	67.9	68.5	66.9	66.9	68	66.9	67.8	67.6%	0.18

Figure 25 shows SPLICE-2 clusters from 10 random initializations.

Node count by Context		Total
227	1039	1266
176	1391	1567
1012	966	1978
670	1241	1911
190	1171	1361
964	971	1935
176	1391	1567
324	1084	1408
285	1232	1517
176	1391	1567
Average		1607.7

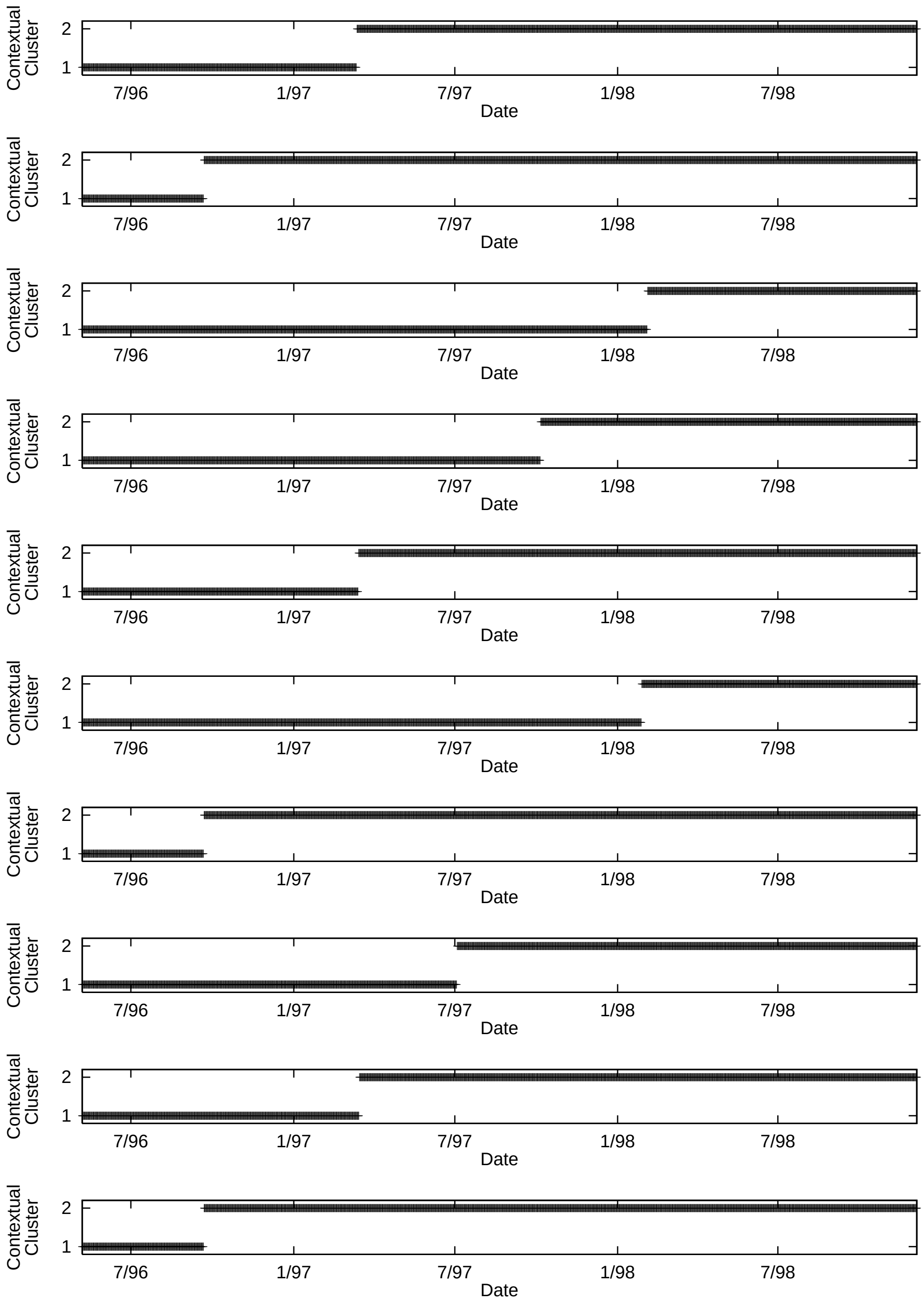

Figure 25: SPLICE-2 clustering on Elec2-3.

A.9 Elec2-1: Three Splice-2 clusters

Dataset	elec2-1
C4.5 arguments	-s
C4.5 walkforward window sizes	336 672 1008 1344 1680 2016 2352 2688 3024 3360 100000
C4.5 walkforward step size	336
Splice-2 arguments	-c4.5arg '-s' -w 2000 -l 10 -r 3
Temporal cross validation settings	-c 336 -k 10

Data

Classes		Default
UP	DOWN	Accuracy
19237	26075	57.5%

Splice-2

Xval Accuracies (%)										Average	Std Err
68.7	67.8	68.9	68.7	68.7	67.8	67.8	68.7	67.8	69.2	68.4%	0.16

Figure 26 shows Splice-2 clusters from 10 random initializations.

Node count by Context			Total
47	24	66	137
44	71	24	139
43	24	65	132
42	60	24	126
24	42	64	130
39	24	57	120
55	42	24	121
24	42	55	121
27	37	78	142
46	24	62	132
Average			130.0

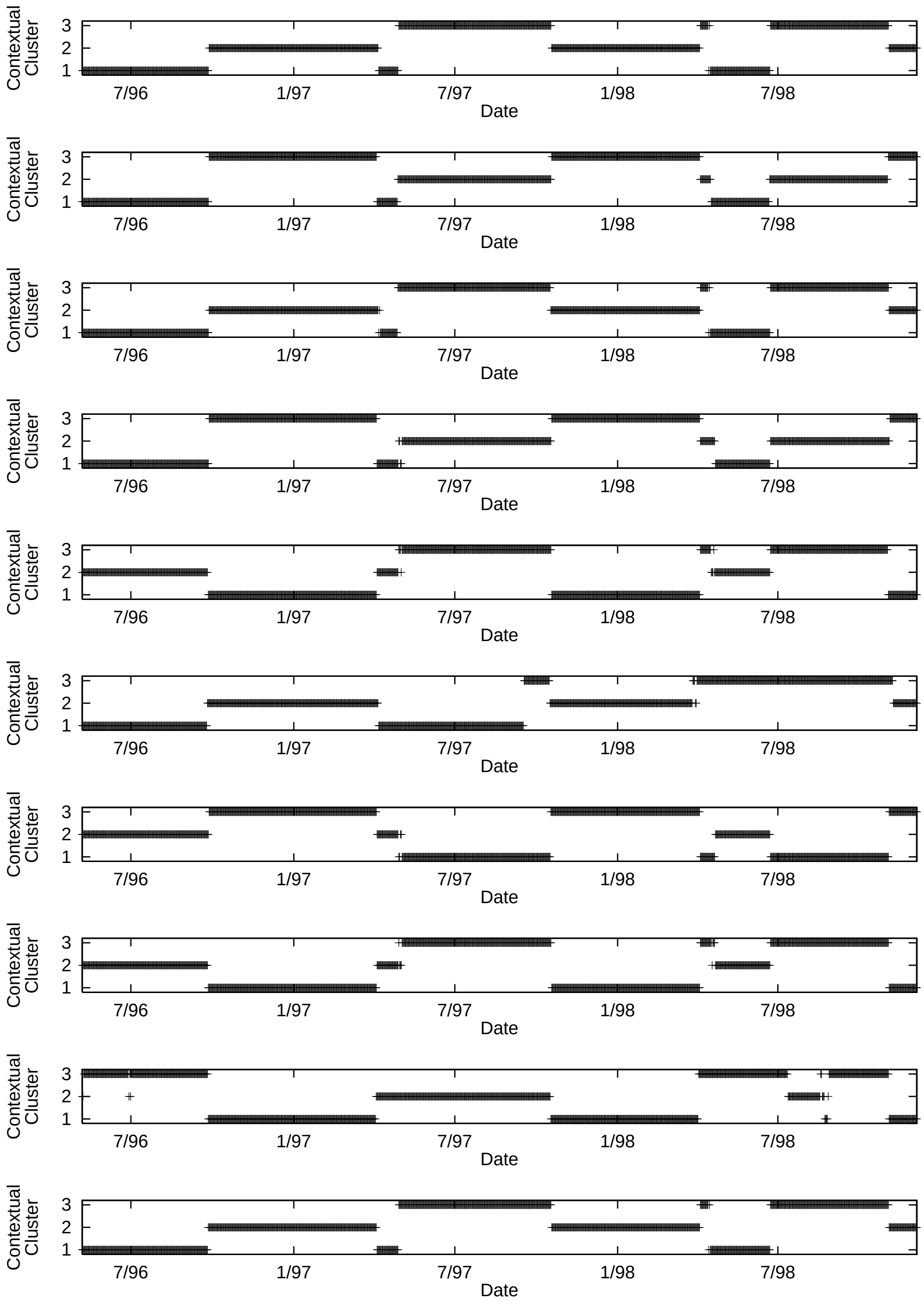

Figure 26: SPLICE-2 clustering on Elec2-1.

A.10 Elec2-2: Three Splice-2 Clusters

Dataset	elec2-2
C4.5 arguments	-s
C4.5 walkforward window sizes	336 672 1008 1344 1680 2016 2352 2688 3024 3360 100000
C4.5 walkforward step size	336
Splice-2 arguments	-c4.5arg '-s' -w 2000 -l 10 -r 3
Temporal cross validation settings	-c 336 -k 10

Data

Classes		Default
UP	DOWN	Accuracy
19237	26075	57.5%

Splice-2

Xval Accuracies (%)										Average	Std Err
68.5	67.3	68.3	69.1	69.2	67.3	67.3	68.7	67.3	69.1	68.2%	0.25

Figure 27 shows Splice-2 clusters from 10 random initializations.

Node count by Context			Total
261	126	236	623
194	213	164	571
165	256	265	686
167	250	281	698
274	128	229	631
230	172	244	646
218	318		536
194	119	253	566
247	160	188	595
200	122	287	609
Average			616.1

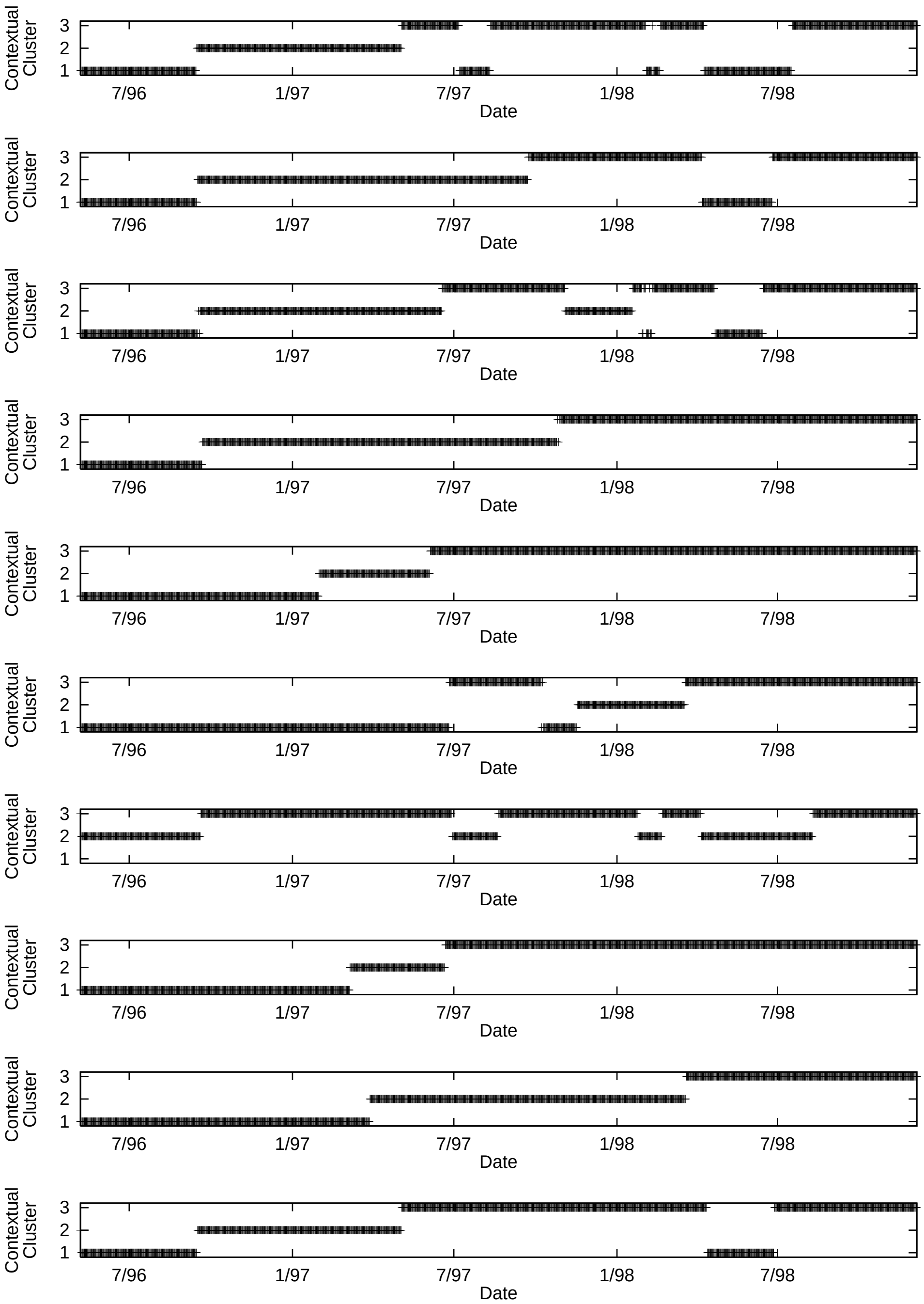

Figure 27: SPLICE-2 clustering on Elec2-2.

A.11 Elec2-3: Three Splice-2 Clusters

Dataset	elec2-3
C4.5 arguments	-s
C4.5 walkforward window sizes	336 672 1008 1344 1680 2016 2352 2688 3024 3360 100000
C4.5 walkforward step size	336
Splice-2 arguments	-c4.5arg '-s' -w 2000 -l 10 -r 3
Temporal cross validation settings	-c 336 -k 10

Data

Classes		Default
UP	DOWN	Accuracy
19237	26075	57.5%

Splice-2

Xval Accuracies (%)										Average	Std Err
68.5	67	68.1	68.7	69	67	67	68.7	67	68.1	67.9%	0.25

Figure 28 shows Splice-2 clusters from 10 random initializations.

Node count by Context			Total
155	156	1180	1491
161	502	1319	1982
169	964	935	2068
161	552	1149	1862
274	167	1074	1515
671	435	1045	2151
176	1391		1567
272	397	1183	1852
285	1043	902	2230
161	99	1202	1462
Average			1818.0

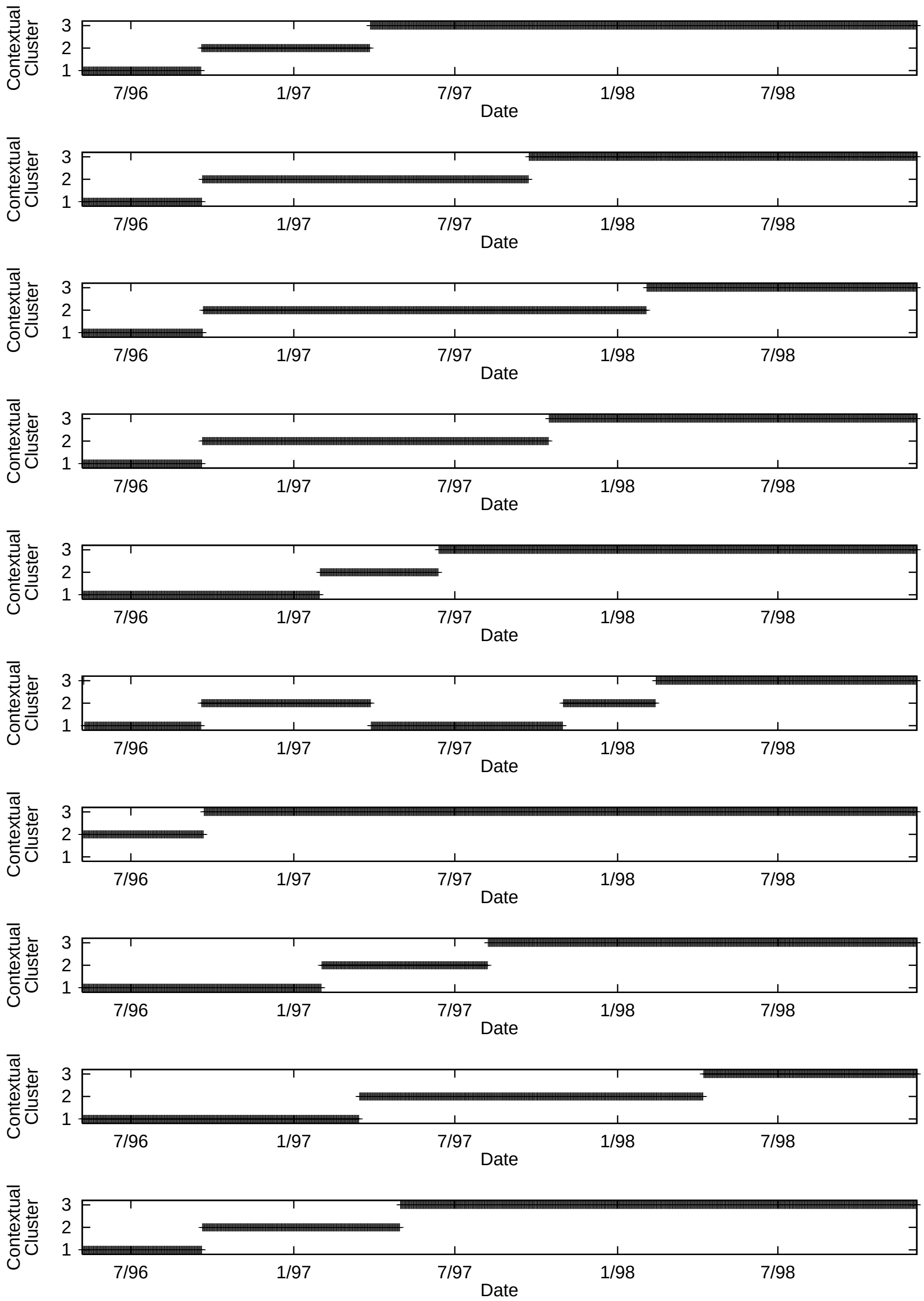

Figure 28: SPLICE-2 clustering on Elec2-3.