Under consideration for publication in Knowledge and Information Systems

Mining Sequential Patterns for Classification

Dmitriy Fradkin[1], Fabian Mörchen[2]

[1] Siemens Corporate Technology, Princeton, NJ
[2] Amazon, Seattle, WA

Keywords: Sequential Pattern Mining, Sequence Classification, Information Gain

Abstract. While a number of efficient sequential pattern mining algorithms were developed over the years, they can still take a long time and produce a huge number of patterns, many of which are redundant. These properties are especially frustrating when the goal of pattern mining is to find patterns for use as features in classification problems. In this paper we describe BIDE-Discriminative, a modification of BIDE that uses class information for direct mining of predictive sequential patterns. We then perform an extensive evaluation on 9 real-life datasets of the different ways in which the basic BIDE-Discriminative can be used in real multi-class classification problems, including 1-vs-rest and model-based search tree approaches. The results of our experiments show that 1-vs-rest provides an efficient solution with good classification performance.

1. Introduction

Temporal data mining exploits temporal information in data sources in the context of data mining tasks such as clustering or classification. Many scientific and business data sources are dynamic and thus promising candidates for application of temporal mining methods. For an overview of methods to mine time series, sequence, and streaming data see (Dong and Pei, 2007),(Han and Kamber, 2006).

One particular type of temporal data consists of sequences of (sets of) discrete items associated with time stamps. Examples of such data include medical records (Batal et al., 2011),(Batal et al., 2012), histories of transactions of customers in an online shop, log messages emitted by machines or telecommunication equipment during operation (Xu et al., 2008),(Sipos et al., 2014) and discretized or abstracted time series or sensor data. A common task is to mine for local regularities in this data by looking for sequential patterns (Agrawal and Srikant, 1995) that represent a sequence of itemsets, possibly

Received May 15, 2014
Revised Oct 08, 2014
Accepted Dec 06, 2014

with gaps, embedded in the observation sequences. Another common task is sequence classification - given a set of example sequences belonging to two or more classes, to learn a predictive model capable of correctly assigning class labels to previously unseen sequences. In this paper we:

- Propose BIDE-Discriminative - a modification to the highly efficient BIDE algorithm (Wang and Han, 2004) for direct predictive pattern mining of sequential data that relies on pruning statistical measures of feature predictiveness.
- Propose approaches (**BIDE-D** and **BIDE-DC**) for using BIDE-Discriminative for direct predictive pattern mining in multi-class problems.
- Propose a combination of BIDE-Discriminative and model-based tree (MbT) classifier (Fan, Zhang, Cheng, Gao, Yan, Han, Yu and Verscheure, 2008) for creating a sequence tree-based classifier (**SMBT**) or for an alternative direct pattern mining approach (**SMBT-FS**).
- Conduct an extensive experimental evaluation of all of the proposed methods against regular BIDE as well as an obvious alternative of separately mining patterns for each class (**BIDE-C**). Our experiments show that the proposed methods are faster and produce fewer patterns than indirect methods, and lead to at least comparable accuracy. They also indicate possible trade-offs in mining speed and accuracy between different methods.

2. Related work

Two types of approaches for classification on itemset and symbolic sequential data have been proposed. These can be thought of as *indirect* and *direct* approaches (Cheng et al., 2008). The indirect approaches consist of three stages. First, candidate pattern are mined in an unsupervised fashion or separately for each class, and then feature selection (Guyon and Elisseeff, 2003) is applied, usually using criteria such as information gain (IG) or χ^2 (Yang and Pedersen, 1997). Finally, a classifier such as SVM is built on the remaining features. Many papers utilize such indirect approaches for classification, ex. (Cheng et al., 2007),(Buza and Schmidt-Thieme, 2010).

The drawbacks of indirect approaches are clear and have been noted in multiple publications (Cheng et al., 2008): (i) generation of all candidate patterns can be computationally very expensive, and (ii) the overwhelming majority of the patterns thus found turn out to be useless for classification or interpretation. Thus, feature selection is still necessary as an additional step before classification to avoid loss in predictive performance. Note that utilizing more restrictive definitions of patterns, such as closed or margin-closed itemset (Lucchese et al., 2006; Moerchen et al., 2010) or sequential (Wang and Han, 2004; Fradkin and Moerchen, 2010) patterns, or reducing the number of patterns by other means (Knobbe and Ho, 2006; Bringmann and Zimmermann, 2008) ameliorates to some extent the first of these problems, but does not address the second.

Direct methods address both of these problems by utilizing class label information in the pattern mining stage, and possibly combining/interleaving the steps of pattern mining and pattern evaluation, thereby generating fewer but better patterns. Furthermore, they can be faster due to pruning of the pattern search space.

2.1. Classification with itemset data

Multiple indirect pattern mining algorithms exist for itemset data, staring with Apriori (Agrawal et al., 1993). Examples include CHARM(Zaki and Hsiao, 2002), FPTree(Han et al., 2000a), FPClose (Grahne and Zhu, 2003). Any of these can be used to generate features for classification.

In (Cheng et al., 2007), a minimum support threshold is set based on information gain bounds for mining discriminative itemsets. FPClose is used to find all itemsets with support greater then this threshold. Maximal Marginal Relevance (Carbonell and Coldstein, 1998) is then applied to further reduce the set of itemsets to be used as features. SVM and decision trees are used as classifiers. This approach can be seen as falling between indirect and direct approaches.

There have also been some real direct approaches. In (Cheng et al., 2008), a branch-and-bound search is performed for discriminative itemset patterns. These patterns are generated sequentially on a shrinking FP-tree (Han et al., 2000b) by eliminating training instances that are sufficiently covered, as a way to reduce redundancy in the set. This requires multiple runs of the mining algorithm to generate one pattern at a time. Model-based search tree (M^bT) approach is proposed in (Fan, Zhang, Cheng, Gao, Yan, Han, Yu and Verscheure, 2008) and evaluated on itemset and on graph datasets. Here a single most discriminative pattern is selected, and the dataset is split into two, based on whether instances have this pattern. Each set is then recursively processed, until all instances in a set belong to the same class, or the set is too small. In this way, a tree for classifying unseen examples in constructed.

None of these approaches have been evaluated on sequential data, which is what we do in the present paper. Our approaches SBMT and SMBT-FS are analogous to those of (Cheng et al., 2008) except for sequential patterns, but we also consider mining top k informative patterns in a single run (BIDE-D and BIDE-DC approaches).

2.2. Classification with sequential data

Just as for itemsets, there are many indirect sequential pattern mining methods. GSP (Srikant and Agrawal, 1996), SPADE(ZAKI, 2001), PrefixSpan(Pei et al., 2001), GSpan (Yan and Han, 2002), BIDE(Wang and Han, 2004) are just some of them. Examples of their use can be found in (Buza and Schmidt-Thieme, 2010) and (Lo et al., 2009). In (Buza and Schmidt-Thieme, 2010) time series are converted into symbolic sequences using Symbolic Aggregate Approximation (Lin et al., 2003). Sequential patterns (motifs) are then mined using GSP/Apriori-like approach with taxonomy. These patterns (motifs) are used as features for construction of SVMs and Bayesian Networks. In (Lo et al., 2009), sequential patterns in software traces are mined using a custom algorithm, then feature selection based on Fisher score is applied and finally a classifier is applied.

We focus below on more direct approaches.

Text and biologic sequences can be considered sequential data, so we mention an approach that was described in (Ifrim et al., 2008),(Ifrim and Wiuf, 2011). Logistic regression models are build in the space of all possible n-grams (of words or symbols) by directly constructing predictive n-grams. Gaps of predefined maximum size (ie. n-grams with wildcards) can also be handled. This approach can therefore be thought of as discovering predictive sequential patterns, though it is quite different from other sequential pattern mining work.

In (Lee et al., 2011), the approach of (Cheng et al., 2007) (discussed above in Section 2.1) was used to set a minimum support threshold based on information gain bound

	Class c_1	...	Class c_k	Total
P present	$s_1(P)$	...	$s_k(P)$	$s(P)$
P absent	$\|c_1\| - s_1(P)$	...	$\|c_k\| - s_k(P)$	$N - s(P)$
	$\|c_1\|$	...	$\|c_k\|$	N

Table 1. Co-occurrence matrix for pattern P and classes C.

for mining sequential patterns. Partial sequential patterns, i.e. of limited length, were mined and then feature selection using F-score was performed before applying classification methods. This approach is thus more of an indirect than a direct method.

In (Lo et al., 2011) dyadic sequential patterns are mined using a BIDE-like (Wang and Han, 2004) algorithm. The instances are pairs of sequences and the task is to determine whether elements in a pair match or not. While the high-level idea is similar to BIDE Discriminative that we propose here, the data, the patterns and the problem are different.

In (Bringmann et al., 2006) multiple pattern languages, including sequences, trees and graphs were compared for their usefulness in a classification task. The patterns were mined using branch-and-bound methods like those described in Section 2.3 with a modification of GSpan (Yan and Han, 2002) (no details provided) and χ^2 as the discriminative measure. In our work we focus on evaluating alternative approaches to finding discriminative sequential patterns, rather than evaluating different pattern languages. Also, we base our discriminative pattern mining algorithm on BIDE, rather than on GSpan.

2.3. Branch-and-bound methods

How can discriminative patterns be discovered without examining all the candidates? The solution, utilized in some form in all of the above-mentioned direct approaches, is to use branch-and-bound search in the space of patterns while maintaining an upper bound on the utility of the search subtree at a pattern.

Pruning based on statistical metrics such as IG or χ^2 was proposed in (Morishita and Sese, 2000) for finding predictive itemsets. An Apriori type algorithm was used to traverse the search space, and upper bound on the feature quality was used for pruning. In (Ohara et al., 2008), such approach was used for discriminative subgraph mining, in (Lo et al., 2011) - for classification of dyadic sequences.

Relation between a pattern's support and discriminative power was discussed in (Cheng et al., 2007), but only in the context of setting appropriate minimum support for frequent itemset pattern mining.

3. Upper Bounds on Discriminative Measures

Suppose we have a dataset D of labeled examples (x_i, y_i), $i = 1, ..., N$ where $y_i \in C$ and C is a set of k class labels: $C = \{c_j\}$, $j = 1, ..., k$. Let $s_j(P)$ be support of pattern P in class c_j - in other words, it is the number of examples of class c_j that contain/match pattern P. Also, let $s(P)$ be the total support of P - the sum of support over all classes $c \in C$. Table 1 shows the relationships between these values.

Information Gain (IG) has long been used for feature selection in classification

(Yang and Pedersen, 1997). For convenience, let $l(p) = p\log_2(p)$. Then IG is defined (for a fixed dataset) as:

$$IG(P) = \sum_{c\in C} l(\frac{|c|}{N}) - \frac{s(P)}{N}\sum_{c\in C} l(\frac{s_c(P)}{s(P)}) \\ -\frac{N-s(P)}{N}\sum_{c\in C} l(\frac{|c|-s_c(P)}{N-s(P)}) \tag{1}$$

Note that the first term in Eq. 1 is fixed for any given dataset, since it depends only on class distribution. Therefore, for a fixed dataset $IG(P)$ is a function of conditional distribution $p(c|x)$ which can be fully represented by a vector of class supports $s_c(P)$. We can write $IG(P) = IG(s_1(P), ..., s_k(P))$.

For a two-class problems, IG is a convex function. Morishita and Sese (Morishita and Sese, 2000) have come up with an upper bound on it:

$$IG_{ub}(r,q) = max(IG(r,0), IG(0,q)) \tag{2}$$

where $r = s_+(P)$ and $q = s_-(P)$.

Analysis in (Morishita and Sese, 2000) shows that for two-class problems other discriminative measures such as chi-square and correlation can also be bounded in a similar fashion. For multi-class problems, bounds (in the same form as above) have been found for chi-squared (Nijssen and Kok, 2006) and inter-class variance (Sese and Morishita, 2004). Both papers used these bounds in analysis of itemset data. Note however that these results apply to any pattern language, such as itemsets, sequential patterns, graphs, etc.

We are not aware of similar upper-bounds for IG in multi-class problems. Theorem 7.4 in (Cover and Thomas, 2006) states that mutual information (aka IG) $I(X, C)$ is a concave function of $p(x)$ for a fixed $p(c|x)$, and a concave function of $p(c|x)$ for a fixed $p(x)$. However, in our case neither one is fixed (extending a pattern changes both), and so IG appears to be neither convex nor concave for more than 2 classes.

Let us state clearly what existence of such upper bounds implies. Given a labelled dataset, patterns P and P' such that P occurs in P' (or P' matches or extends P - we'll define this formally in Section 2.3) and a discriminativeness measure F with a known upper bound, if $F_{ub}(P) = x$, then $F(P') \leq x$. This enables construction of branch-and-bound algorithms for pattern mining, such as those we mentioned in Section 2.3 and those we present in this paper, using any discriminative measure with a defined upper bound. Basically, if we already observe a pattern with discriminative score x, and $F_{ub}(P) = y < x$, then we know that no extension of P is going to have a higher discriminative score, and this part of pattern space does not need to be explored. The discriminative measure itself does not need to be (and usually is not) anti-monotone.

In this paper we focus on IG, while noting that the same methods are applicable to chi-square, correlation and other discriminativeness measures where an upper bound can be formulated.

4. BIDE Family of Algorithms

4.1. Definitions

An **event sequence** over a set of events Σ is a sequence of pairs (t_i, s_i) of **event sets** $s_i \subseteq \Sigma$, $\forall i = 1, \dots, n$ and time stamps $t_i \in \Re^+$. The ordering is based on time, i.e.

$\forall i < j : t_i \leq t_j$. The **length** of the event sequence is n. For the present discussion (and in much of sequential pattern mining literature) the exact values of the time stamps are not as important as the ordering that they impose. We will therefore treat event sequences as just an ordered set of event sets $S = \{s_i\}$.

A **sequence database**, SDB, of size N is a collection of event sequences S_i, $i = 1, \ldots, N$.

A sequence $S = \{s_i\}, i = 1, ..., k$ **matches** a sequential pattern $P = \{p_j\}, j = 1, ..., m$ (or a pattern P **occurs** in the sequence S) iff $\exists i_1, ..., i_m$ with $p_j \subseteq s_{i_j}$ for $j = 1, ..., m$, such that $\forall 1 \leq j, k \leq m$: $p_j \prec p_k$ implies $i_j < i_k$. We will denote such a **match** by $m_{i_1, i_m}(P, S)$. A match $m_{i_1, i_m}(P, S)$ is the **earliest match** iff for any other match $m_{j_1, j_m}(P, S)$ $i_k \leq j_k, \forall k = 1, \ldots, m$,

Support of a pattern, support(P), is the number of sequences where the pattern occurs. A pattern is **frequent** if its support is above some predefined minimum support μ. A pattern is **closed** if none of the patterns that include it have the same support.

In order to describe the algorithms in the following sections, we need to introduce the notion of **projected databases**, which is extremely useful in constructing efficient algorithms for sequential pattern mining.

Given a pattern $P = \{p_i\}$, $i = 1, \ldots, |P|$ and a sequence $S = \{s_j\}$,$j = 1, \ldots, |S|$, with the earliest match $m_{k_1, k_m}(P, S)$, a projection of S on P results in a **projected sequence** $S|P = \{s_t\}$, where $t = k_m + 1, \ldots, |S|$. We refer to k_m as an offset.

Given a pattern $P = \{p_i\}$, $i = 1, \ldots, |P|$ and an SDB $D = \{S_j\}$,$j = 1, \ldots, |D|$, a projection of D on P is a **projected database** $D|P$, consisting of projected sequences $S_j|P$, obtained by projecting S_j onto P. Note that if a sequence does not match a pattern, it does not appear in the projected database. Projected database $D|P$ can be efficiently represented with a list of pairs of indices (j, t), where j refers to $S_j|P$ and t is the corresponding offset.

4.2. BIDE

Details of BIDE implementation can be found in (Wang and Han, 2004),(Wang et al., 2007),(Fradkin and Moerchen, 2010). BIDE is initially called with the full sequential database D, minimum support μ and an empty pattern $P = \emptyset$. It returns a list of frequent closed sequential patterns. BIDE operates by recursively extending patterns, and, while their frequency is above the minimum support, checking closure properties of the extensions.

Consider a frequent pattern $P = \{p_i\}$,$i = 1, \ldots, n$. There are two ways to extend pattern P *forward* with item j:

- Appending the set $p_{n+1} = \{j\}$ to P obtaining $P' = p_1 \ldots p_n p_{n+1}$, called a forward-S(equence)-extension.
- Adding j to the last itemset of P: $P' = p_1 \ldots p'_n$, with $p'_n = p_n \cup j$, assuming $j \notin p_n$, called a forward-I(tem)-extension.

Similarly, a pattern can be extended *backward*

- Inserting the set $p_x = \{j\}$ into P anywhere before the last set obtaining $P' = p_1 \ldots p_i p_x p_{i+1} \ldots p_n$, for some $0 \leq i \leq n$, called a backward-S(et)-extension.
- Adding j to any set in P obtaining $P' = p_1 \ldots p'_i \ldots p_n$, with $p'_i = p_i \cup j$, assuming $j \notin p_n$, $1 \leq i \leq n$, called a backward-I(tem)-extension.

According to Theorem 3 of (Wang et al., 2007), a pattern is closed if there exists no forward-S-extension item, forward-I-extension item, backward-S-extension item, nor backward-I-extension item with the same support.

Furthermore, if there is a backwards extension item, then the resulting extension and all of its future extension are explored in a different branch of the recursion, meaning that it can be pruned from current analysis. These insights are combined in BIDE, leading to a very memory-efficient algorithm, because the patterns found do not need to be kept in memory while the algorithm is running.

Algorithm 1 BIDE Algorithm

```
Require: Sequential Pattern $P = \{p_i\}$, Projected Database $D|P$, minimum support $\mu$
 1: $F$ - set of frequent closed patterns (global variable)
 2: $l = |P|$
 3: $Ls = sStepFrequentItems(P, D|P, \mu)$;
 4: $Li = iStepFrequentItems(P, D|P, \mu)$;
 5: if !$(freqCheck(Ls, P)||freqCheck(Li, P))$ then
 6:     if $backscan(P, D, true)$ then
 7:         $F = F \cup P$
 8: for itemset $p \in Ls$ do
 9:     $P' = p_1, .., p_l, p$
10:     if $backscan(P', D|P', false)$ then
11:         $bide(P', D|P', \mu)$;
12: for itemset $p \in Li$ do
13:     $P' = p_1, .., p_{l-1}, p_l \cup p$
14:     if $backscan(P', D|P', false)$ then
15:         $bide(P', D|P', \mu)$;
16: return $F$
```

Specifically, consider pseudo-code for BIDE (Algorithm 1). In Lines 3-4 items that can be used in forward extension of the current pattern are found. If there is no forward extension with the same support (Line 5), the backward closure is checked (Line 6) using function backscan. If the pattern is also backwards-closed, it can be added to the set of closed frequent patterns (Line 7).

Then, we check every item in forward S and I extensions (in the two for-loops) to see whether it is explored in a different branch of recursion, again via backscan function (Lines 10 and 14). If not, then we project the database on the extension and call BIDE recursively on the extension and the new projected database. We will omit additional details as they are not relevant to our discussion.

4.3. BIDE-Discriminative

In order to select only discriminative patterns, we need to keep track of the class label information. We assume that each sequence in dataset D is associated with a class label, and thus we know class distribution in the dataset. When we search for potential S and I extensions, we also keep track of class distribution of sequences where they occur. This gives us all the necessary information for determining discriminative scores and their upper bounds for any new pattern that we discover. Let $d(P, c)$ denote discriminative score for pattern P for class c, and let $d_{UB}(P, c)$ denote the upper bound on discriminative score for any extension P'' of P for class c. When discussing two-class problems, c can be omitted from the notation.

Algorithm 2 BIDE Discriminative Algorithm

```
Require: Sequential Pattern $P = \{p_i\}$, Projected Database $D|P$, minimum support $\mu$,
    $k$ - number of patterns to be selected
 1: $F$ - set of discriminative patterns (global variable)
 2: $dt = 0$ - minimal threshold for discriminative score of a pattern (global variable)
 3: if $d_{UB}(P) < dt$ then
 4:    return
 5: $l = |P|$
 6: $Ls = sStepFrequentItems(P, D|P, \mu)$;
 7: $Li = iStepFrequentItems(P, D|P, \mu)$;
 8: if $d(P) \geq dt$ then
 9:    if $!(freqCheck(Ls, P)||freqCheck(Li, P))$ then
10:       if $backscan(P, D, true)$ then
11:          $F = F \cup P$
12:          if $|F| > k$ then
13:             $F = F - argmin_{X \in F} d(X)$
14:             $dt = \min_{X \in F} d(X)$
15: for itemset $p \in Ls$ do
16:    $P' = p_1, .., p_l, p$
17:    if $backscan(P', D|P', false)$ then
18:       $BIDEDiscriminative(P', D|P', \mu)$;
19: for itemset $p \in Li$ do
20:    $P' = p_1, .., p_{l-1}, p_l \cup p$
21:    if $backscan(P', D|P', false)$ then
22:       $BIDEDiscriminative(P', D|P', \mu)$;
23: return $F$
```

The pseudo-code for BIDE-Discriminative is given in Algorithm 2. We have a new user-specified parameter k - the number of patterns to extract. We introduce variable dt, a threshold on discriminative score, which is initially set to 0.

In Lines 3-4 we check the upper bound of pattern P. If the upper bound is below the threshold dt, then the pattern and all of its extensions can be pruned - the function returns. Otherwise, algorithm proceeds. In Line 8, the score of the pattern itself is checked against the threshold, to determine if it should be added to the list. If the size of the set F exceeds k, the pattern with the lowest score is removed, and threshold dt is updated accordingly (Lines 12-14). Regardless of whether pattern itself is added to the set, its extensions are still examined like in regular BIDE, since the upper bound was above the threshold.

Note that with minor modifications we can have BIDE-Discriminative output all patterns with discriminative score above some value. All that is needed is to remove parameter k and make dt a fixed parameter instead.

4.4. Handling Multi-Class Problems

As mentioned before, the upper bounds for IG or χ^2 hold for two-class problems only. Multi-class problems can be handled in two ways.

BIDE-D: Run BIDE-Discriminative on the whole training data once, but redefine discriminative scores and upper bounds of patterns to be maximums over all binary one-vs-rest problems. Specifically:

$$d'(P) = max_{c \in C} d(P, c) \tag{3}$$

$$d'_{UB}(P) = max_{c \in C} d_{UB}(P, c) \tag{4}$$

BIDE-DC: Mine discriminative patterns on the whole training data separately for each of $|C|$ one-vs-rest binary problems. (This straightforward idea has been mentioned in (Nijssen and Kok, 2006) in context of itemset pattern mining, but not implemented). The sets of top-k patterns produced for each class can be merged, resulting in at most $k|C|$ patterns. Note that this could be done in a single run, by maintaining separate sets of patterns and thresholds for each class and expanding a pattern as long as it could be informative for at least one class (**BIDE-DC-1R**). Alternatively, the same set of patterns can be discovered by performing $|C|$ one-vs-rest runs of two-class discriminative BIDE (**BIDE-DC-CR**).

These direct approaches can be contrasted with two indirect approaches:

BIDE: Run regular BIDE on all training data and use all closed patterns found. This approach is fully unsupervised.

BIDE-C: Run BIDE separately on training data from each class and merge the sets of patterns found. Note that this is the only approach of the four that does mining on subsets of the whole dataset. This approach is indirect but does make some use of class labels.

4.5. Computational Efficiency

A single run of BIDE-Discriminative has the same complexity as BIDE, since it may still potentially generate all frequent closed sequential patterns, in exactly the same fashion. However, the pruning, depending on the values of k or dt, may remove a significant number of patterns and their extensions from consideration. The cost involved in computing the upper bounds for each pattern is independent of the size of the data (at most proportional to number of classes) with appropriate bookkeeping, and so does not affect computational complexity.

The reduction in the number of patterns produced can also lead to savings in I/O time and space/memory to store them. Furthermore, since the discriminative scores for the generated patterns are already known, there is no need to separately compute them them, as in indirect approaches.

The two variants of BIDE-DC allow us to compare the costs of deeper expansion and of maintaining multiple pattern sets, against the costs of doing $|C|$ runs of BIDE-Discriminative.

Meanwhile, BIDE-C requires $|C|$ runs of BIDE, but each is only on a fraction of the dataset, which is likely to be much faster than a single run on the full dataset. We explore these trade-offs in the experimental section.

5. Model-Based Tree Algorithms

A method for direct construction of a tree-based predictive model (model-based tree or MbT) was proposed in (Fan, Zhang, Cheng, Gao, Yan, Han, Yu and Verscheure, 2008) and evaluated on itemset and on graph data. We combine BIDE-Discriminative algorithm with the MbT idea for efficiently building a predictive model for sequential

Algorithm 3 BIDE-DC-1R: Discriminative Algorithm for Multiclass Problems

```
Require: Sequential Pattern $P = \{p_i\}$, Projected Database $D|P$, minimum support $\mu$,
    $k$ - number of patterns to be selected per class
$F_c$ - a set of sets of discriminative patterns for each class (global variable)
$dt_c = 0$ - a vector of minimal thresholds for discriminative scores of a pattern for
    each class (global variable)
if $\forall c \in C$: $d_{UB}(P, c) < dt_c$ then
    return
$l = |P|$
$Ls = sStepFrequentItems(P, D|P, \mu)$;
$Li = iStepFrequentItems(P, D|P, \mu)$;
if $!(freqCheck(Ls, P)||freqCheck(Li, P))$ then
    if $backscan(P, D, true)$ then
        for class $c \in C$ do
            if $d(P, c) \geq dt_c$ then
                $F_c = F_c \cup P$
                if $|F_c| > k$ then
                    $F_c = F_c - argmin_{X \in F} d(X, c)$
                    $dt_c = \min_{X \in F_c} d(X, c)$
for itemset $p \in Ls$ do
    $P' = p_1, .., p_l, p$
    if $backscan(P', D|P', false)$ then
        $BIDE - DC - 1R(P', D|P', \mu)$;
for itemset $p \in Li$ do
    $P' = p_1, .., p_{l-1}, p_l \cup p$
    if $backscan(P', D|P', false)$ then
        $BIDE - DC - 1R(P', D|P', \mu)$;
return $\cup F_c$
```

data. We will refer to this as **SMBT**. The pseudocode is given in Algorithm 4. BIDE-Discriminative (Line 4) is used to find the most informative pattern. The database is then split into two sets of sequences: those that match this pattern, and those that don't. The two sets are processed recursively in the same fashion, until they become too small or until the class purity of a node exceeds a predefined threshold.

Algorithm 4 SMBT: MBT Algorithm for sequential data

```
Require: Projected Database $D$, minimum support $\mu$, maximum purity $p$, minimum
    leaf size $z$
1: if $(|D| < z)||(purity(D) > p)$ then
2:     Create and return a leaf node $N$ with label of majority class of $D$
3: Create an empty tree node $N$
4: $P = BIDEDiscriminative(D, \mu, k = 1)$
5: $N.P = P$ - assign pattern $P$ to the node
6: $N.left = MBT(S|P \notin S, \mu)$
7: $N.right = MBT(S|P \in S, \mu)$
8: return $N$
```

The resulting tree can be used to make prediction on test data in a straightforward fashion. Starting with the root node, we check if a node is a leaf (in which case we return

Data	Intervals	Labels	Sequences	Classes
ASL-BU (Papaterou et al., 2005)	18250	154	441	7
ASL-GT (Starner et al., 1998)	89247	47	3493	40
Auslan2 (Kadous, 2002)	900	12	200	10
Blocks (Fern, 2004)	1207	8	210	8
Context (Mäntyjärvi et al., 2004)	12916	54	240	5
Pioneer (Asuncion and Newman, n.d.)	4883	92	160	3
Skating (Mörchen and Ultsch, 2007)	18953	41	530	6 (7)
Unix (Asuncion and Newman, n.d.)	147504	1598	7762	9
WW3D (Kerr et al., 2008)	11597	107	157	6

Table 2. Datasets used in experiments

the class label for the node). If the node is not a leaf, we check if the node pattern occurs in the sequence and depending on the result descent to the left or to the right child.

We expect the SMBT to include only a small fraction of the patterns that would have been produced by BIDE-Discriminative alone, significantly easing interpretation of the classifier and of the extracted patterns.

While SMBT approach directly constructs a classifier, it can also be viewed as a form of feature selection (**SMBT-FS**): the patterns in the SMBT can be treated as individual features, to be combined using methods such as support vector machines or neural nets.

6. Experiments

So far, we have described the following approaches:

- two indirect (i.e. unsupervised) sequential pattern mining approaches (BIDE and BIDE-C)
- two variants of BIDE for directly mining discriminative patterns, specifically: BIDE-D and BIDE-DC (with subvariants BIDE-DC-1R and BIDE-DC-CR)
- a sequential model-based tree approach, SMBT, which is a direct mining method and a complete classifier in itself
- SMBT-FS, which uses SMBT purely as a Feature Selection / direct mining step and builds a classifier such as SVM with the patterns that SMBT extracts as features

The goals of our experiments are to examine and compare the behavior of the above methods, under different parameter choices, in terms of (i) accuracy, (ii) number of patterns produced, and (iii) time required to mine the patterns. Our expectations are that direct mining methods produce fewer pattern and require less time than indirect and unsupervised methods without sacrificing accuracy. Specifically, because of their respective designs, we expect BIDE-D and BIDE-DC to produce fewer patterns and run faster than BIDE, while SMBT will be faster still with even fewer patterns. The behavior of BIDE-C is more difficult to predict.

We use Information Gain as the discriminativeness measures with all of these algorithms.

6.1. Data

While unlabeled sequential data is relatively common, ex. web log dataset (Asuncion and Newman, n.d.), labeled data needed for our evaluation is much harder to come by. The nine datasets used in our experiments are summarized in Table 2. Two (Unix and WW3D) are simple sequential datasets. The remaining seven, while technically databases of intervals, can be interpreted as sequential databases by treating start and end boundaries of an interval as separate events (Wu and Chen, 2007). Specifically, each symbolic interval, a triple (t_s, t_e, σ) with event $\sigma \in \Sigma$ and time stamps $t_s \leq t_e$, is converted into two symbolic time points (t_s, σ^+) and (t_e, σ^-), and then all time points with the same time stamp are aggregated into itemsets, resulting in a standard event sequence.

The advantage of this collection is that class labels are available for each sequence. This allows an automated evaluation of patterns using a classifier, while the categorical sequential data available in the UCI Machine Learning Repository (Asuncion and Newman, n.d.), such as web log data, is largely unlabeled.

ASL-BU[1] The intervals are transcriptions from videos of American Sign Language expressions provided by Boston University (Papaterou et al., 2005). It consists of observation interval sequences with labels such as *head mvmt: nod rapid* or *shoulders forward* that belong to one of 7 classes like *yes-no question* or *rhetorical question*.

ASL-GT The intervals are derived from 16 dimensional numerical time series with features derived from videos of American Sign Language expressions (Starner et al., 1998). The numerical time series were discretized into 2-4 states each using Persist (Mörchen and Ultsch, 2005). Each sequence represents one of 40 word like *brown* or *fish*.

Auslan2 The intervals were derived from the high quality Australian Sign Language dataset in the UCI repository (Asuncion and Newman, n.d.) donated by Kadous (Kadous, 2002). The x,y,z dimensions were discretized using Persist with 2 bins, 5 dimensions representing the fingers were discretized into 2 bins using the median as the divider. Each sequence represents a word like *girl* or *right*.

Blocks[2] The intervals describe visual primitives obtained from videos of a human hand stacking colored blocks provided by (Fern, 2004). The interval labels describe which blocks touch and the actions of the hand (*contacts blue red*, *attached hand red*). Each sequence represents one of 8 different scenarios from atomic actions (*pick-up*) to complete scenarios (*assemble*).

Context[3] The intervals were derived from categoric and numeric data describing the context of a mobile device carried by humans in different situations (Mäntyjärvi et al., 2004). Numeric sensors were discretized using 2-3 bins chosen manually based on exploratory data analysis. Each sequence represents one of five scenarios such as *street* or *meeting*.

Pioneer The intervals were derived from the Pioneer-1 datasets in the UCI repository (Asuncion and Newman, n.d.). The numerical time series were discretized into 2-4 bins by choosing thresholds manually based on exploratory data analysis. Each sequence describes one of three scenarios: *gripper*, *move*, *turn*.

Skating The intervals were derived from 14 dimensional numerical time series describing muscle activity and leg position of 6 professional In-Line Speed Skaters during controlled tests at 7 different speeds on a treadmill (Mörchen and Ultsch, 2007). The

[1] `http://www.bu.edu/asllrp/`

[2] `ftp://ftp.ecn.purdue.edu/qobi/ama.tar.Z`

[3] `http://www.cis.hut.fi/jhimberg/contextdata/index.shtml`

time series were discretized into 2-3 bins using Persist and manually chosen thresholds. Each sequence represents a complete movement cycle and is labeled by skater or speed.

Unix The dataset consists of sanitized command histories of 9 users (Asuncion and Newman, n.d.).

WW3D This dataset was collected from Wubble World 3D (ww3d), a virtual environment with simulated physics in which softbots, called wubbles, interact with objects (Kerr et al., 2008).

6.2. Experiment Setup

For each method and parameter setting we repeated 5-fold cross-validation 3 times, each time with a different random split. Thus all the measures (accuracy, number of patterns, run times, etc) are averages taken over 3*5=15 test sets. Each experimental run consisted of two parts: pattern mining and classification. Pattern mining is done on the training folds. Then sequences in both training and test folds are converted to binary vectors based on presence of the patterns. These vectors are written to files. (This part is programmed in Java). The vectors from training folds are then used to build classifiers, which are then applied to the test folds.

For direct BIDE approaches (BIDE-D and both variants of BIDE-DC), we considered $k = 10, 20, 30, 40, 50, 70, 90$, where k is the number of patterns per class.

In order to compare all the approaches, we also had to find a way of setting minimum support similarly for all of them. For most datasets we used the following strategy: parameter $\nu = 0.2, 0.3, 0.4, 0.5, 0.6$ specified the fraction of the size of the smallest class in the training data that would be used as minimum support. For example, if the training set has 100 instances of 5 classes, but the smallest class has 10 instances, then with $\nu = 0.5$ the minimum support would be set to 5 for all methods.

It turned out that for some datasets (ASL-GT, Context, Skating, Unix) this approach didn't work - pattern mining was taking too long for all the approaches. For these datasets we changed the way minimum support was computed: fraction ν of the whole training set size was used. BIDE-C cannot be used in such situations, but the other three methods ran successfully. In these cases, ν is equivalent to μ, minimum support. For Context and Skating $\nu = 0.7, 0.8, 0.9$. For ASL-GT dataset, $\nu = 0.2, 0.3, 0.4, 0.5, 0.6$; while for Unix dataset $\nu = 0.1, 0.15, 0.2$.

The way we set minimum support ensures that all four methods are exploring the same pattern space, and could conceivably produce exactly the same set of patterns - i.e. a set of closed frequent patterns for minimum support derived using ν. It follows that the set of patterns found by BIDE is going to be a superset of sets of patterns found by the other approaches for the same value of ν.

For SMBT experiments, we kept the same values of ν. However we had to additionally specify values for minimum leaf size z and leaf purity p. We experimented with $p = 0.8, 0.9$ and $z = 1, 3, 5$.

Classification was performed in Matlab, using LIBLINEAR (Fan, Chang, Hsieh, Wang and Lin, 2008) (a fast implementation of linear SVM) with options ”-B 1 -S 5”, i.e. L1-regularized L2-loss support vector classification. The value for parameter C was selected from the set $10^{-3,\cdots,3}$ using 3-fold cross-validation on the training set.

6.3. Experimental Results

Here we describe the highlights of our evaluation. More details are presented in the appendix[4].

Comparison of Two BIDE-DC variants: The two variants, BIDE-DC-1R and BIDE-DC-CR will produce identical sets of patterns, so the only meaningful comparison between them is in speed. Not surprisingly, using a single run (BIDE-DC-1R) is always significantly faster than doing separate runs (BIDE-DC-CR), despite some additional overhead and deeper search required for the former. (These results are shown in Appendix).Thus, in the rest of the paper we will use only BIDE-DC-1R, and for simplicity will refer to it as BIDE-DC.

Effect of k: We experimented with $k = 10, 20, 30, 40, 50, 70, 90$, as mentioned in Section 6.2. Performance of BIDE-DC seems to be less sensitive to value of k and of ν than that of BIDE-D, likely because it distributes patterns evenly across classes, while BIDE-D may suffer from redundant patterns that are all predictive for the same class. Higher values of k lead to better results where the differences are observed. Thus in the rest of the experiments, we will keep k fixed at 90. (The plots showing accuracy vs. ν for different values of k are shown in Appendix).

Effects of purity and leaf size: SMBT and SMBT-FS are not particularly sensitive to choices of minimum leaf size and purity (see Appendix) Thus we focus on results with leaf size 5 and purity of 0.9.

Accuracy Comparisons: We start by comparing the accuracies of the most similar approaches: BIDE vs BIDE-C, BIDE-D vs BIDE-DC, and SMBT vs SMBT-FS. We then compare performance of 'winners' to each other. The Appendix shows complete results and discusses them in more detail.

BIDE leads to better results than BIDE-C on ASL-BU and Blocks, and worse ones on Auslan2 dataset. On the other datasets however BIDE-C is difficult to apply due to the need to set minimum support ν separately for each class. Thus BIDE is to be generally preferred between these two. (Figure 1a).

BIDE-DC outperforms BIDE-D on ASL-BU, Context, Pioneer and WW3D, while having comparable accuracy on the other datasets. (Figure 1b).

SMBT-FS noticeably outperforms SMBT on ASL-BU, and Unix, is slightly worse on Auslan2, Context and WW3D, and comparable on the other datasets. As can be seen in Appendix, SMBT-FS also performs better against other methods (i.e. SMBT does better on datasets where both methods perform poorly). Thus SMBT-FS is a better method.

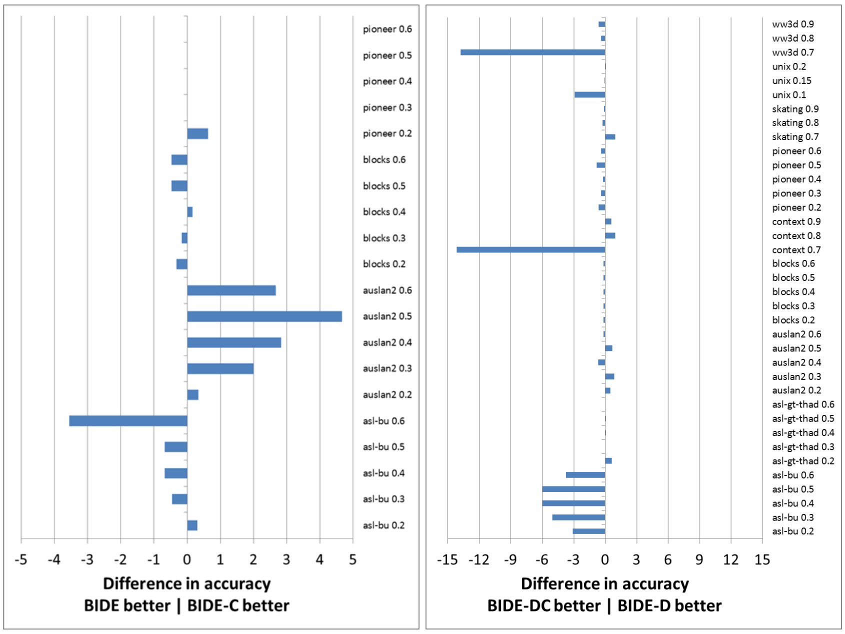
We now compare the 'winners' from each pair. Comparison between BIDE-DC and SMBT-FS favors the former, which performs much better on most datasets, only slightly worse on Skating and Auslan2 (Figure 2b). SMBT-FS also is clearly worse than BIDE, again excepting Skating and Auslan2 (Figure 3a). BIDE-DC and BIDE are closely matched (Figure 3b), with the differences in accuracy less than 2% in either direction in almost all the cases.

We can thus conclude that BIDE-DC is the best and most stable of the direct methods in terms of accuracy, and gives performance comparable to using all the patterns mined in an unsupervised fashion, i.e. BIDE. SMBT-FS can occasionally perform somewhat better than BIDE-DC, but also frequently produces much worse results.

Number of Patterns and Mining Speed:

Having compared the accuracies of different approaches, we turn our attention to

[4] https://sites.google.com/site/dfradkin/kais2014-separateAppendix.pdf



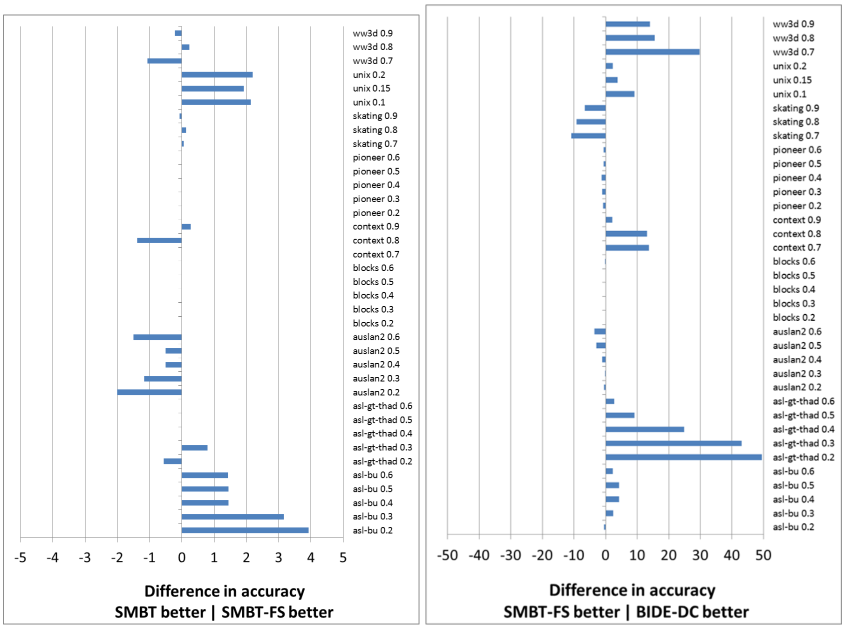
(a) (b)

Fig. 1. Comparison of accuracy of BIDE approaches.

the number of patterns mined and computational efficiency. Figure 4a shows, on log scale, the ratio of patterns produced by BIDE to that produced by BIDE-DC. It is obvious that on most datasets BIDE-DC uses just a small fraction of frequent closed patterns that BIDE generates. Figure 4b shows reduction in pattern mining time from using BIDE-DC, compared to using BIDE. BIDE-DC is faster, sometimes by a lot, except on ASL-GT and Auslan2 datasets. Note that these are the same datasets where the number of patterns produced by BIDE is comparable/same as that produced by BIDE. The explanation for these results is that on some datasets, for certain values of ν few patterns exist, and BIDE finds them all faster than BIDE-DC due to smaller overhead. However, when the number of potential patterns is high, BIDE-DC is faster.

Finally, we demonstrate that using only the top discriminative features leads to improved training time for a classifier, such as SVM. Figure 5a compares training times with patterns produced by BIDE and BIDE-DC, with latter consistently faster except on ASL-GT and Auslan, as discussed.

We also note that SMBT-FS can sometime produce a lot fewer patterns, and be several times faster in mining time and training time (Figure 5b), than BIDE-DC, but again, the cost in terms of accuracy can be high. Detailed results are presented in the Appendix.



(a) (b)

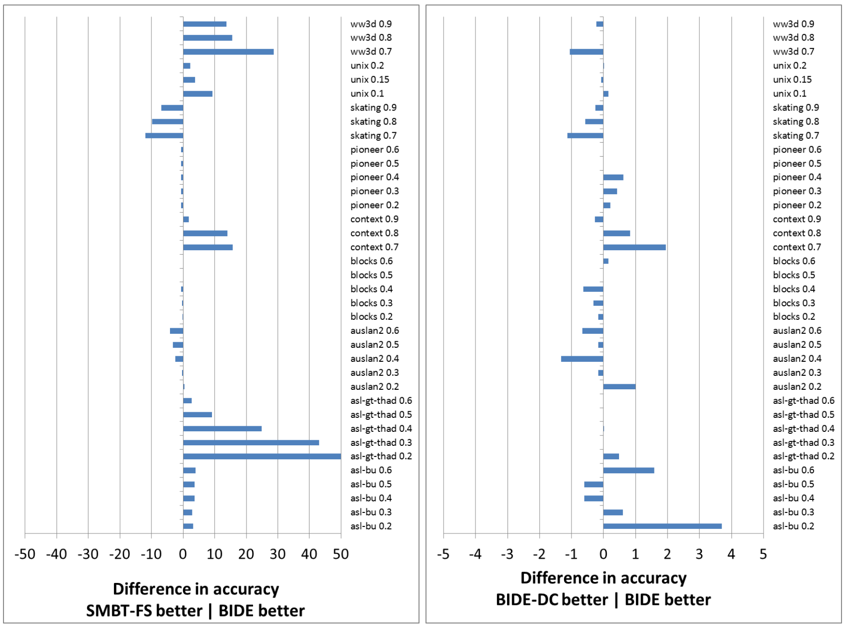
Fig. 2. Comparison of accuracy of SMBT-FS with SMBT and BIDE-DC.

7. Conclusions

We have described a direct sequential pattern mining approach, BIDE-Discriminative, and how it can be utilized for discriminative sequential pattern mining in multi-class problems. We have evaluated approaches for discriminative sequential pattern mining in multi-class problems (BIDE-D, BIDE-DC, SMBT and SMBT-FS) against unsupervised approaches (BIDE and BIDE-C). Our experiments suggest that BIDE-DC is usually the best option, as it efficiently generates a small number of predictive patterns leading to comparable classification performance while potentially saving the user order of magnitude in memory, and noticeable amount of time both during the pattern mining stage and when training a classifier. The slight advantage in accuracy obtained by BIDE-DC over BIDE-D is likely due to former extracting fewer redundant patterns.

SMBT and SMBT-FS can sometimes match accuracy of BIDE-DC with a lot fewer patterns, but can also result in significantly worse performance. Also, while in some cases they were faster than the other methods, in others the reverse was true. This unpredictability makes us caution against these approaches, though in some application domains where very small models are required the benefits may outweigh the risks.

Our evaluation was performed on a collection of 9 real-world datasets, and is the first evaluation of sequential discriminative pattern mining methods.



(a) (b)

Fig. 3. Comparison of accuracy of SMBT-FS, BIDE-DC and BIDE.

References

Agrawal, R., Imielinski, T. and Swami, A. N. (1993), Mining association rules between sets of items in large databases, *in* 'Proc. of the 1993 ACM SIGMOD Intl. Conf. on Management of data', ACM Press, pp. 207–216.

Agrawal, R. and Srikant, R. (1995), Mining sequential patterns, *in* 'ICDE', IEEE Press, pp. 3–14.

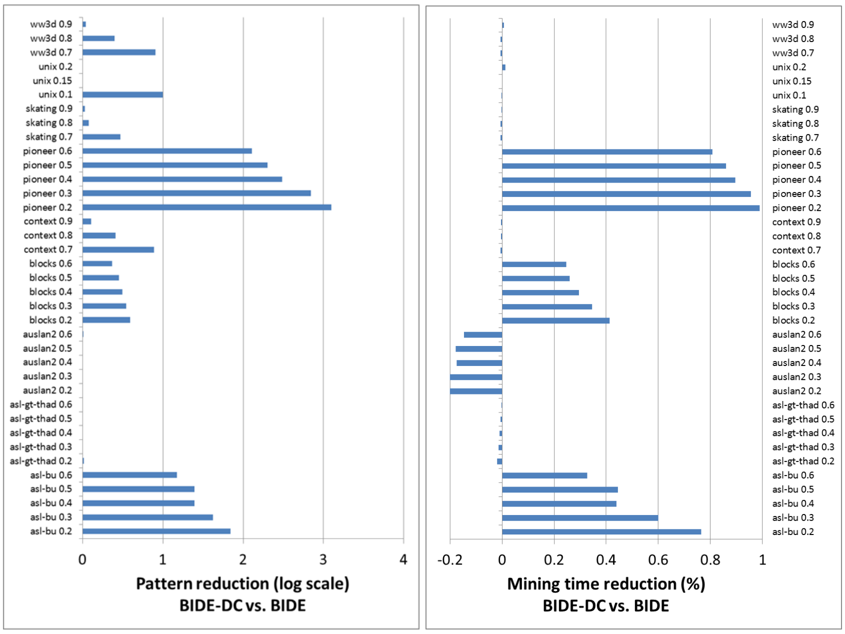
Asuncion, A. and Newman, D. (n.d.), 'UCI Machine Learning Repository'.

Batal, I., Fradkin, D., Harrison, J., Moerchen, F. and Hauskrecht, M. (2012), Mining recent temporal patterns for event detection in multivariate time series data, *in* 'Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', pp. 280–288.
URL:`http://doi.acm.org/10.1145/2339530.2339578`

Batal, I., Valizadegan, H., Cooper, G. F. and Hauskrecht, M. (2011), A pattern mining approach for classifying multivariate temporal data, *in* 'Proceedings of the 2011 IEEE International Conference on Bioinformatics and Biomedicine', pp. 358–365.
URL:`http://dx.doi.org/10.1109/BIBM.2011.39`

Bringmann, B. and Zimmermann, A. (2008), 'One in a million: picking the right patterns', *KAIS* **18**(1), 61–81.

Bringmann, B., Zimmermann, A., Raedt, L. and Nijssen, S. (2006), Dont be afraid of simpler patterns, *in* J. Frnkranz, T. Scheffer and M. Spiliopoulou, eds, 'Knowledge Discovery in Databases: PKDD 2006', Vol. 4213 of *LNCS*, Springer Berlin Heidel-



(a) (b)

Fig. 4. Comparison of number of patterns and mining times of BIDE-DC and BIDE.

berg, pp. 55–66.
URL:http://dx.doi.org/10.1007/11871637_10

Buza, K. and Schmidt-Thieme, L. (2010), Motif-based classification of time series with bayesian networks and svms, *in* A. Fink, B. Lausen, W. Seidel and A. Ultsch, eds, 'Advances in Data Analysis, Data Handling and Business Intelligence', Studies in Classification, Data Analysis, and Knowledge Organization, Springer Berlin Heidelberg, pp. 105–114.
URL:http://dx.doi.org/10.1007/978-3-642-01044-6_9

Carbonell, J. and Coldstein, J. (1998), The use of mmr, diversity-based reranking for reordering documents and producing summaries, *in* 'In Proc. of SIGIR', p. 335336.

Cheng, H., Yan, X., Han, J. and Hsu, C.-W. (2007), Discriminative frequent pattern analysis for effective classification, *in* 'Proc. IEEE ICDE'.

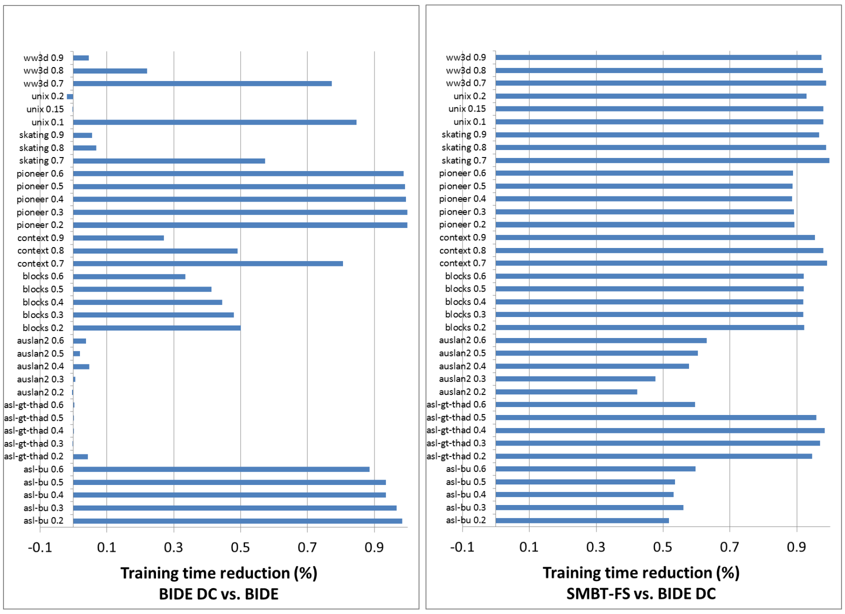
Cheng, H., Yan, X., Han, J. and Yu, P. S. (2008), Direct discriminative pattern mining for effective classification, *in* 'ICDE', pp. 169–178.

Cover, T. M. and Thomas, J. A. (2006), *Elements of information theory*, 2 edn, Wiley.

Dong, G. and Pei, J. (2007), *Sequence Data Mining*, Morgan Kaufmann.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. and Lin, C.-J. (2008), 'Liblinear: A library for large linear classification', *Journal of Machine Learning Research* **9**, 1871–1874.

Fan, W., Zhang, K., Cheng, H., Gao, J., Yan, X., Han, J., Yu, P. S. and Verscheure, O.



(a) (b)

Fig. 5. Comparison of training times of SMBT-FS, BIDE-DC and BIDE.

(2008), Direct mining of discriminative and essential frequent patterns via model-based search tree, *in* 'KDD', pp. 230–238.

Fern, A. (2004), Learning Models and Formulas of a Temporal Event Logic, PhD thesis, Purdue University, West Lafayette, IN, USA.

Fradkin, D. and Moerchen, F. (2010), Margin-closed frequent sequential pattern mining, *in* 'KDD Workshop on Useful Patterns', ACM, New York, NY, USA, pp. 45–54.

Grahne, G. and Zhu, J. (2003), Efficiently using prefix-trees in mining frequent itemsets, *in* 'ICDM Workshop on Frequent Itemset Mining Implementations'.

Guyon, I. and Elisseeff, A. (2003), 'An introduction to variable and feature selection', *JMLR* **3**, 1157–1182.

Han, J. and Kamber, M. (2006), *Data Mining - Concepts and Techniques, 2nd edition*, Morgan Kaufmann.

Han, J., Pei, J. and Yin, Y. (2000*a*), Mining frequent patterns without candidate generation, *in* 'Proc. ACM SIGMOD Intl. Conf. on Management of Data', ACM Press, pp. 1–12.

Han, J., Pei, J. and Yin, Y. (2000*b*), Mining frequent patterns without candidate generation, *in* 'SIGMOD', pp. 1–12.

Ifrim, G., Bakir, G. H. and Weikum, G. (2008), Fast logistic regression for text categorization with variable-length n-grams., *in* 'KDD', pp. 354–362.

Ifrim, G. and Wiuf, C. (2011), 'Bounded coordinate-descent for biological sequence classification in high dimensional predictor space', *KDD* .

Kadous, M. W. (2002), Temporal Classification: Extending the Classification Paradigm to Multivariate Time Series, PhD thesis, University of New South Wales.

Kerr, W., Cohen, P. and Chang, Y.-H. (2008), Learning and playing in wubble world, *in* 'Proceedings of the Fourth Artificial Intelligence and Interactive Digital Entertainment Conference', pp. 66–71.

Knobbe, A. J. and Ho, E. K. Y. (2006), Pattern teams, *in* 'PKDD', pp. 577–584.

Lee, J.-G., Han, J., Li, X. and Cheng, H. (2011), 'Mining discriminative patterns for classifying trajectories on road networks', *IEEE Transactions on Knowledge and Data Engineering* **23**(5), 713–726.

Lin, J., Keogh, E., Lonardi, S. and Chiu, B. (2003), A symbolic representation of time series, with implications for streaming algorithms, *in* 'Proc. of the 2003 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery', ACM Press, pp. 2–11.
URL:`http://citeseer.ist.psu.edu/583097.html`

Lo, D., Cheng, H. and Cia, L. (2011), Mining closed discriminative dyadic sequential patterns, *in* 'EDBT'.

Lo, D., Han, J., Cheng, H., Khoo, S.-C. and Sun, C. (2009), Classification of software behaviros for failure detection: A discriminative pattern mining approach, *in* 'Proceedings of KDD'.

Lucchese, C., Orlando, S. and Perego, R. (2006), 'Fast and memory efficient mining of frequent closed itemsets', *IEEE TKDE* **18**(1), 21–36.

Mäntyjärvi, J., Himberg, J., Kangas, P., Tuomela, U. and Huuskonen, P. (2004), Sensor signal data set for exploring context recognition of mobile devices, *in* 'Proc. of PERVASIVE', Springer, pp. 18–23.

Moerchen, F., Thies, M. and Ultsch, A. (2010), 'Efficient mining of all margin-closed itemsets with applications in temporal knowledge discovery and classification by compression.', *Knowledge and Information Systems* .

Mörchen, F. and Ultsch, A. (2005), Optimizing time series discretization for knowledge discovery, *in* 'Proc. ACM SIGKDD', ACM Press, pp. 660–665.

Mörchen, F. and Ultsch, A. (2007), 'Efficient mining of understandable patterns from multivariate interval time series', *Data Min. Knowl. Discov.* .

Morishita, S. and Sese, J. (2000), Traversing itemset lattice with statistical metric pruning, *in* 'PODS', pp. 226–236.

Nijssen, S. and Kok, J. (2006), Multi-class correlated pattern mining, *in* F. Bonchi and J.-F. Boulicaut, eds, 'Knowledge Discovery in Inductive Databases', Vol. 3933 of *LNCS*, Springer Berlin Heidelberg, pp. 165–187.
URL:`http://dx.doi.org/10.1007/11733492_10`

Ohara, K., Hara, M., Takabayashi, K., Motoda, H. and Washio, T. (2008), Pruning strategies based on the upper bound of information gain for discriminative subgraph mining., *in* 'PKAW'08', pp. 50–60.

Papaterou, P., Kollios, G., Sclaroff, S. and Gunopoulos, D. (2005), Discovering frequent arrangements of temporal intervals, *in* 'ICDM', pp. 354–361.

Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U. and Hsu, M.-C. (2001), PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth, *in* 'Proc. IEEE ICDE', IEEE Press, pp. 215–224.

Sese, J. and Morishita, S. (2004), Itemset classified clustering, *in* J.-F. Boulicaut, F. Esposito, F. Giannotti and D. Pedreschi, eds, 'Knowledge Discovery in Databases:

PKDD 2004', Vol. 3202 of *LNCS*, Springer Berlin Heidelberg, pp. 398–409. URL:http://dx.doi.org/10.1007/978-3-540-30116-5_37

Sipos, R., Fradkin, D., Moerchen, F. and Wang, Z. (2014), Log-based predictive maintenance, *in* 'Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', pp. 1867–1876. URL:http://doi.acm.org/10.1145/2623330.2623340

Srikant, R. and Agrawal, R. (1996), Mining sequential patterns: Generalizations and performance improvements, *in* 'Proc. of the 5th Intl. Conf. on Extending Database Technology (EDBT)', Springer, pp. 3–17. URL:http://citeseer.ist.psu.edu/article/srikant96mining.html

Starner, T., Weaver, J. and Pentland, A. (1998), 'Real-time American Sign Language recognition using desk and wearable computer-based video', *IEEE TPAMI* **20**(12).

Wang, J. and Han, J. (2004), BIDE: Efficient mining of frequent closed sequences, *in* 'ICDE', IEEE Press, pp. 79–90.

Wang, J., Han, J. and Li, C. (2007), 'Frequent closed sequence mining without candidate maintenance', *IEEE TKDE* **19**(8), 1042–1056.

Wu, S.-Y. and Chen, Y.-L. (2007), 'Mining nonambiguous temporal patterns for interval-based events', *IEEE TKDE* **19**(6), 742–758.

Xu, W., Huang, L., Fox, A., Patterson, D. and Jordan, M. (2008), Mining console logs for large-scale system problem detection., *in* 'Proceedings of the 3rd Workshop on Tackling Computer Systems Problems with Machine Learning Techniques'.

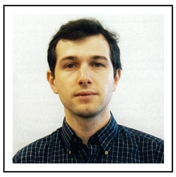
Yan, X. and Han, J. (2002), gspan: Graph-based substructure pattern mining, *in* 'ICDM'.

Yang, Y. and Pedersen, J. (1997), A comparative study on feature selection in text categorization, *in* 'ICML', pp. 412–420.

ZAKI, M. (2001), 'Spade: An efficient algorithm for mining frequent sequences', *Machine Learning* **42**, 31–60.

Zaki, M. J. and Hsiao, C.-J. (2002), CHARM: An efficient algorithm for closed itemset mining, *in* 'Proc. of the 2nd SIAM Intl. Conf. on Data Mining (SDM)', SIAM, pp. 457–473.

Author Biographies



Dmitriy Fradkin is a Senior Scientist at Siemens Corporate Technology, Princeton NJ. He received B.A. in Mathematics and Computer Science from Brandeis University, Waltham, MA in 1999 and Ph.D. from Rutgers, The State University of New Jersey in 2006. Before joining Siemens in 2007 he has worked at Ask.com. His research is in applying data mining and machine learning techniques to solve real-world problems is areas of predictive maintenance, healthcare and text analytics. Dr. Fradkin is a member of the ACM SIGKDD, and a reviewer for several data mining journals.

Fabian Mörchen graduated with a Ph.D. in 2006 from the Philipps University of Marburg, Germany with summa cum laude. His thesis contributed novel methods in mining temporal pattern from interval data. From 2006-2012 worked at Siemens Corporate Research leading data mining projects with applications in predictive maintenance, text mining, healthcare, and sustainable energy. He continued his research in temporal data mining in the context of industrial and scientific problems and served the community as a reviewer, organizer of workshops, and presenter of tutorials. Since 2012 he is leading a data science team at Amazon to improve customer experience using machine learning and big data analytics.

Correspondence and offprint requests to: Dmitriy Fradkin, Siement Corporate Technology, 755 College Rd East, Princeton, NJ 08540 USA. Email: dmitriy.fradkin@siemens.com