Information Engineering Express
International Institute of Applied Informatics
2019, Vol. 5, No. 1, 86 – 96

Iterative Consistency-Based Feature Selection and Its Application to Nucleotide Sequences of Influenza A Viruses

Sho Shimamura [*], Kouichi Hirata [*]

Abstract

In this paper, first we formulate a consistency-based feature selection problem as combinatorial optimization problems. Next, for the purpose of increasing the number of instances explained by the features, which we call *explanatory instances*, rather than decreasing the number of features themselves in consistency-based feature selection, we introduce an *iterative consistency-based feature selection* and design the algorithm to compute it. Finally, we apply the method to several nucleotide sequences of influenza A viruses and evaluate the advantage of the method.

Keywords: Iterative Consistency-Based Feature Selection, Consistency-Based Feature Selection, CWC, LCC, Nucleotide Sequences, Influenza A Viruses.

1 Introduction

It is one of important social problems to characterize influenza viruses to predict next trend influenza viruses. Then, it is effective to analyze nucleotide sequences of influenza viruses from the viewpoint of bioinformatics or medical informatics.

In general, it is a standard method to analyze nucleotide sequences of influenza viruses by using a well-known *alignment*. As another methods, Makino *et al.* [9] have introduced a *trim distance* between positions (or sites) in nucleotide sequences based on phylogenetic trees reconstructed from nucleotide sequences. Also Shimada *et al.* [14] have investigated the *clustering* by the trim distance to analyze 2009 pandemic of influenza A (H1N1) viruses. Furthermore, Hamada *et al.* [7] have applied several kernels including an *agreement sub-tree mapping kernel* for phylogenetic trees reconstructed from nucleotide sequences for influenza A viruses to 2009 pandemic classification and regional analysis. Whereas these researches have suggested some characterization of sites, they have the problem to reconstruct phylogenetic trees.

In order to avoid this problem and characterize the relative sites in nucleotide sequences of influenza viruses directly, Shimamura and Hirata [15] have regarded the sites as features and first applied *feature selection* [5][6][10] to nucleotide sequences. Then, they have analyzed temporal and regional characters of nucleotide sequences of influenza viruses by using *consistency-based feature selection* algorithms [3][8][11][22].

[*] Kyushu Institute of Technology, Iizuka, Japan

An LCC (Linear Consistency-Constrained) [18] and a CWC (Combination of Weakest Components) [19][20][21], which this paper mainly deals with, are known as fast and accurate consistency-based feature selection algorithms based on a *Bayesian risk* and a *binary consistency*, respectively, as consistency measures. Both algorithms are greedy backward elimination algorithms excluding features. On the other hand, there arises a problem for both algorithms that, when increasing the number of instances explained by the features, which we call *explanatory instances*, rather than decreasing the number of features themselves, they may eliminate too many inconsistent instances in general.

Concerned with this problem, the consistency-based feature selection essentially contains two directions of optimizations, the minimization of the number of features and the maximization of the number of instances explained by the features. Then, in this paper, we first formulate it as combinatorial optimization problems.

Next, in order to increase the number of explanatory instances, in this paper, we introduce an *iterative consistency-based feature selection*, by applying the feature selection algorithm to the eliminated instances in a data set iteratively. In other words, we design the method to obtain the "disjunction" of feature sets iteratively. This method is possible to be more effective for data with many features such as a nucleotide sequence whose number of features is its length.

Hence, in this paper, we apply the iterative consistency-based feature selection to nucleotide sequences of influenza A viruses as the previous work [15]. Here, we deal with nucleotide sequences of influenza A viruses for 4 subtypes of H1N2, H2N2, H3N2 and N5N1 and 8 RNA segments of PB2, PB1, PA, HA, NP, NA, MP and NS. All of the nucleotide sequences are provided from NCBI [4]. Then, we observe that the number of explanatory instances for 3 subtypes of H1N2, H3N2 and N5N1 and all the 8 RNA segments is always increasing by iterative consistency-based feature selection of CWC, and that for all the 4 subtypes and the RNA segment PA is always increasing by iterative consistency-based feature selection of both LCC and CWC.

This paper is organized as follows. In Section 2, we formulate a consistency-based feature selection problem as combinatorial optimization problems. In Section 3, we introduce the algorithms LCC and CWC and design the algorithm of iterative consistency-based feature selection. In Section 4, we give experimental results by applying the iterative consistency-based feature selection to nucleotide sequences. Section 5 concludes this paper.

2 Consistency-Based Feature Selection

In this paper, we formulate a consistency-based feature selection by using a matrix on natural numbers. Then, we regard a feature as the set of the numbers of columns except the last column and the last column as the class labels.

We call an $m \times (n+1)$ matrix on $\mathbf{N}$ a *data set* and denote it by $D = [v_{ij}]$. Also we call every row $v_i = [v_{i1}, \dots, v_{in}, v_{i(n+1)}]$ in D an *instance* of D and the $(n+1)$-th element $v_{i(n+1)}$ in v_i a *class label* of v_i. We denote the set of all the class labels in D by C. In the following, we omit the subscript i. Then, we denote that v is an instance of D by $v \in D$ and the class label of v by v_c.

Let $F = \{1, \dots, n\}$, which we call a *total feature set*, and $v = v_i \in D$ an instance. Then, we denote $[v_{i1}, \dots, v_{in}]$ by v_F. For a subset $X = \{j_1, \dots, j_k\} \subseteq F$, which we call a *feature set*, we denote $[v_{ij_1}, \dots, v_{ij_k}]$ by v_X. For a data set D and a feature set $X \subseteq F$, we denote the data set consisting of the j-th column for every $j \in X \cup \{n+1\}$, that is, the collection of rows $[v_X, v_c]$ for $v \in D$ by D_X.

In this paper, we deal with two consistency measures, a *Bayesian risk* [18] and a *binary consistency* [19][20][21]. For $X \subseteq F$, the *Bayesian risk* $BR(X)$ of X is defined as follows:

$$BR(X) = 1 - \sum_{X \subseteq F} \max_{y \in C} Pr(D_X = [v_X, y]).$$

On the other hand, the *binary consistency* $BC(X)$ supports:

$$BC(X) = \begin{cases} 0 & \forall u, v \in D(u_X = v_X \Rightarrow u_c = v_c), \\ 1 & \text{otherwise.} \end{cases}$$

Let $\mu \in \{BR, BC\}$ be a consistency measure and δ a threshold ($0 \leq \delta < 1$). Then, we say that X is *consistent with respect to* D *under* μ and δ if $\mu(X) \leq \delta$; *inconsistent* otherwise. Note here that δ is not necessary when $\mu = BC$.

Let D be a data set, $X \subseteq F$ a feature set, $\mu \in \{BR, BC\}$ a consistency measure and δ a threshold. Then, we call the set of instances by eliminating all the inconsistent instances of D for X under μ and δ from D the set of *explanatory instances* of D for X and denote it by $e_{\mu,\delta}(D,X)$. It is obvious that X is consistent with $e_{\mu,\delta}(D,X)$ under μ and δ. Then, we formulate *a consistency-based feature selection problem* (CONFS) as follows.

> CONFS (*cf.*, [18][19][20][21])
> INSTANCE: A data set D, a total feature set F, a consistency measure μ and a threshold δ.
> PROBLEM: Find a feature set $X \subseteq F$ such that $|X|$ is minimum when $|e_{\mu,\delta}(D,X)|$ is maximum.

As same as standard feature selection problems, the problem CONFS is intractable, because the problem of finding X such that $|e_{\mu,\delta}(D,X)|$ is maximum for the same input is at least NP-hard (*cf.*, [1][2]).

3 Iterative Consistency-Based Feature Selection Algorithms

In order to solve the problem CONFS heuristically and efficiently, Shin *et al.* have introduced the algorithms LCC (Linear Consistency-Constrained) [18] and CWC (Combination of Weakest Components) [19][20][21] illustrated in Algorithm 1. Here, the procedure **sort** sorts F as $\{i_1, \dots i_n\}$ by increasing order of *symmetric uncertainty* [12], which is a normalized value of mutual information [13] of C and $X \subseteq F$ and denote by $SU(C,X)$. Also the procedure **denoise** removes presumable noise examples from D.

Note that the number of explanatory instances of D is monotonic w.r.t. feature sets, that is, $|e_{\mu,\delta}(D,X)| \leq |e_{\mu,\delta}(D,Y)|$ for $X \subseteq Y \subseteq F$. Also it holds that $SU(C,X)$ is maximum when $|e_{\mu,\delta}(D,X)|$ is maximum and $SU(C,X) \geq \sum_{i \in X} SU(C,\{i\})$.

By using these properties, for the problem CONFS, we can regard that the algorithm CWC finds a feature set $X \subseteq F$ such that $\sum_{i \in X} SU(C,\{i\})$ is maximum and $|X|$ is minimum when $|e_{\mu,\delta}(D,X)|$ is maximum. Also we can regard that the algorithm LCC finds a feature set $X \subseteq F$ such that $\sum_{i \in X} SU(C,\{i\})$ is maximum and $|X|$ is minimum when $|e_{\mu,\delta}(D,X)|/|D| \geq |e_{\mu,\delta}(D,F)|/|D| - \delta$ holds.

In this paper, we design the algorithm ITFS in Algorithm 2 for iterative consistency-based feature selection. Here, FS(D,F) returns the result of the algorithm FS (which is

```
procedure LCC$_\delta$($D$,$F$)
    /* $D$: data set, $F$: total feature set, $\delta$: threshold */
    sort $F$ as $\{i_1, \ldots i_n\}$ by increasing order of symmetric uncertainty;
    $S \leftarrow \{i_1, \ldots, i_n\}$;
    for $j = 1$ to $n$ do
        if $BR(S \setminus \{i_j\}) \leq \delta$ then
            $S \leftarrow S \setminus \{i_j\}$;
    output $S$;
procedure CWC($D$,$F$)
    /* $D$: data set, $F$: total feature set */
    denoise $D$;
    sort $F$ as $\{i_1, \ldots i_n\}$ by increasing order of symmetric uncertainty;
    $S \leftarrow \{i_1, \ldots, i_n\}$;
    for $j = 1$ to $n$ do
        if $BC(S \setminus \{i_j\}) = 0$ then
            $S \leftarrow S \setminus \{i_j\}$;
    output $S$;
```

Algorithm 1: LCC and CWC.

```
procedure ItFS(FS,$D$,$F$)
      /* FS: feature selection algorithm, $D$: data set, $F$: total feature set */
      /* In FS: $\mu \in \{BR, BC\}$: consistency measure, $\delta$: threshold */
  1   $X \leftarrow \emptyset$; $Y \leftarrow \emptyset$;
  2   repeat
  3       $Y \leftarrow$ FS($D$,$F$); $X \leftarrow X \cup Y$;
  4       $D' \leftarrow e_{\mu,\delta}(D, Y)$;
  5       $D \leftarrow D \setminus D'$; $F \leftarrow F \setminus X$;
  6   until $D' = \emptyset$ ;
  7   output $X$;
```

Algorithm 2: ItFS.

either an LCC$_\delta$ or a CWC in this paper) for a current data set D and a current total feature set F. The output of ItFS is a feature set X.

The algorithm ItFS first returns a feature set Y as the result of FS(D,F) and updates a feature set X as $X \cup Y$ in the line 3. Then, it finds the set D' of explanatory instances in the line 4. Finally, it updates D as $D \setminus D'$ and F as $F \setminus X$ in the line 5. The algorithm ItFS repeats the above procedures until $D' = \emptyset$.

4 Experimental Results

In this section, we apply the algorithm ITFS to nucleotide sequences of influenza A viruses for 4 subtypes of H1N1, H2N2, H3N2 and N5N1 and 8 RNA segments of PB2, PB1, PA, HA, NP, NA, MP and NS, provided from NCBI [4].

Tables 1, 2, 3, 4, 5, 6, 7 and 8 illustrate the results of applying the algorithm ItFS to nucleotide sequences of 8 segments for influenza A viruses with the 4 subtypes. Here, we set the threshold δ in the algorithm LCC to 0 and we denote LCC$_0$ as a feature selection algorithm FS by LCC simply.

In their tables, m is the number of instances and n is the number of total features for every segment of nucleotide sequences. Also $|e|$ is the cardinality $|q_{\mu,\delta}(D,X)|$ of explanatory instances, where μ and δ are determined by FS, and $|e|/m$ is the ratio (%) of $|e|$ for m. Furthermore, $|X|$ is the number of selected features by the algorithms FS and ITFS from the data sets, where FS is either an LCC or a CWC and $|X|/n$ is the ratio (%) of $|X|$ for n Finally, # is the number of iterations in the algorithm ITFS. Note that the value of m for H2N2 is much smaller than other segments.

Table 1: The results for the segment of PB2.

subt.	FS	m	n	results of FS				results of ITFS				
				$\|e\|$	$\|e\|/m$	$\|X\|$	$\|X\|/n$	#	$\|e\|$	$\|e\|/m$	$\|X\|$	$\|X\|/n$
H1N1	LCC	11772	2725	9053	76.90	900	33.03	2	9053	76.90	1133	41.58
	CWC			8854	75.21	777	28.51	7	9011	76.55	1019	37.39
H2N2	LCC	189	2341	187	98.94	33	1.41	-	-	-	-	-
	CWC			187	98.94	33	1.41	-	-	-	-	-
H3N2	LCC	11604	2438	8895	76.65	2184	89.58	2	8895	76.65	2329	95.53
	CWC			8666	74.68	755	30.97	8	8863	76.38	1060	43.48
H5N1	LCC	2858	2426	2771	96.96	196	8.08	2	2771	96.96	222	9.15
	CWC			2762	96.64	189	7.79	5	2765	96.75	233	9.60

Table 2: The results for the segment of PB1.

subt.	FS	m	n	results of FS				results of ITFS				
				$\|e\|$	$\|e\|/m$	$\|X\|$	$\|X\|/n$	#	$\|e\|$	$\|e\|/m$	$\|X\|$	$\|X\|/n$
H1N1	LCC	11641	2650	8886	76.33	927	34.98	-	-	-	-	-
	CWC			8673	74.50	792	29.89	10	8865	76.15	1100	41.51
H2N2	LCC	189	2341	183	96.83	41	1.75	2	183	96.83	45	1.92
	CWC			182	96.30	41	1.75	-	-	-	-	-
H3N2	LCC	11618	3185	8841	76.10	901	28.29	-	-	-	-	-
	CWC			8646	74.42	768	24.11	6	8822	75.93	1048	32.90
H5N1	LCC	2901	2520	2778	95.76	204	8.10	2	2778	95.76	240	9.52
	CWC			2741	94.48	182	7.22	5	2750	94.79	231	9.17

Table 3: The results for the segment of PA.

subt.	FS	m	n	results of FS				results of ItFS				
				$\|e\|$	$\|e\|/m$	$\|X\|$	$\|X\|/n$	#	$\|e\|$	$\|e\|/m$	$\|X\|$	$\|X\|/n$
H1N1	LCC	11782	3817	8972	76.15	927	24.29	2	8972	76.15	1035	27.12
	CWC			8803	74.72	807	21.14	7	8922	75.73	1012	26.51
H2N2	LCC	186	2233	178	95.70	44	1.97	2	178	95.70	48	2.15
	CWC			175	94.09	42	1.88	2	175	94.09	44	1.97
H3N2	LCC	11572	2660	8589	74.22	915	34.40	2	8589	74.22	1132	42.56
	CWC			8391	72.51	790	29.70	8	8577	74.12	1066	40.08
H5N1	LCC	2823	2347	2732	96.78	197	8.39	2	2732	96.78	225	9.59
	CWC			2714	96.14	185	7.88	2	2716	96.21	201	8.56

Table 4: The results for the segment of HA.

subt.	FS	m	n	results of FS				results of ItFS				
				$\|e\|$	$\|e\|/m$	$\|X\|$	$\|X\|/n$	#	$\|e\|$	$\|e\|/m$	$\|X\|$	$\|X\|/n$
H1N1	LCC	29114	2358	23333	80.14	1252	53.10	-	-	-	-	-
	CWC			23140	79.48	1004	42.58	9	23305	80.05	1359	57.63
H2N2	LCC	250	1779	246	98.40	46	2.59	2	246	98.40	48	2.70
	CWC			244	97.60	44	2.47	-	-	-	-	-
H3N2	LCC	29441	2253	22691	77.07	1121	49.76	-	-	-	-	-
	CWC			22507	76.45	1007	44.70	7	22649	76.93	1308	58.06
H5N1	LCC	5900	2061	5606	95.02	2061	100.00	-	-	-	-	-
	CWC			5556	94.17	283	13.73	9	5588	94.71	439	21.30

Table 5: The results for the segment of NP.

subt.	FS	m	n	results of FS				results of ItFS				
				$\|e\|$	$\|e\|/m$	$\|X\|$	$\|X\|/n$	#	$\|e\|$	$\|e\|/m$	$\|X\|$	$\|X\|/n$
H1N1	LCC	12345	1636	7872	63.77	805	49.21	-	-	-	-	-
	CWC			7724	62.57	623	38.08	5	7828	63.41	789	48.23
H2N2	LCC	190	1566	184	96.84	48	3.07	2	184	96.84	50	3.19
	CWC			181	95.26	47	3.00	2	181	95.26	49	3.13
H3N2	LCC	12366	1843	7993	64.64	688	37.33	-	-	-	-	-
	CWC			7874	63.67	602	32.66	6	7972	64.47	749	40.64
H5N1	LCC	2937	1631	2795	95.17	203	12.45	2	2795	95.17	242	14.84
	CWC			2767	94.21	185	11.34	9	2785	94.82	287	17.60

Table 6: The results for the segment of NA.

subt.	FS	m	n	results of FS				results of ItFS				
				$\|e\|$	$\|e\|/m$	$\|X\|$	$\|X\|/n$	#	$\|e\|$	$\|e\|/m$	$\|X\|$	$\|X\|/n$
H1N1	LCC	23779	1775	17218	72.41	998	56.23	-	-	-	-	-
	CWC			17108	71.95	865	48.73	5	17196	72.32	1053	59.32
H2N2	LCC	231	1469	231	100.00	42	2.86	-	-	-	-	
	CWC			231	100.00	42	2.86	-	-	-	-	-
H3N2	LCC	17079	1694	12619	73.89	851	50.24	2	12619	73.89	943	55.67
	CWC			12474	73.04	764	45.10	5	12596	73.75	942	55.61
H5N1	LCC	4344	1832	4076	93.83	285	15.56	2	4076	93.83	360	19.65
	CWC			4026	92.68	255	13.92	7	4052	93.28	387	21.12

Table 7: The results for the segment of MP.

subt.	FS	m	n	results of FS				results of ItFS				
				$\|e\|$	$\|e\|/m$	$\|X\|$	$\|X\|/n$	#	$\|e\|$	$\|e\|/m$	$\|X\|$	$\|X\|/n$
H1N1	LCC	17407	1223	8537	49.04	621	50.78	-	-	-	-	-
	CWC			8437	48.47	562	45.95	8	8504	48.85	717	58.63
H2N2	LCC	217	1028	193	88.94	1028	100.00	-	-	-	-	-
	CWC			185	85.25	46	4.47	3	188	86.64	59	5.74
H3N2	LCC	16648	1123	8361	50.22	538	47.91	-	-	-	-	-
	CWC			8301	49.86	498	44.35	4	8347	50.14	597	53.16
H5N1	LCC	3273	1138	2919	89.18	552	48.51	-	-	-	-	-
	CWC			2854	87.20	202	17.75	10	2905	88.76	355	31.20

Table 8: The results for the segment of NS.

subt.	FS	m	n	results of FS				results of ItFS				
				$\|e\|$	$\|e\|/m$	$\|X\|$	$\|X\|/n$	#	$\|e\|$	$\|e\|/m$	$\|X\|$	$\|X\|/n$
H1N1	LCC	12684	992	7068	55.72	648	65.32	-	-	-	-	-
	CWC			6971	54.96	570	57.46	6	7053	55.61	682	68.75
H2N2	LCC	201	891	187	93.03	39	4.38	2	187	93.03	44	4.94
	CWC			176	87.56	34	3.82	2	176	87.56	40	4.49
H3N2	LCC	12355	928	6243	50.53	928	100.00	-	-	-	-	-
	CWC			6172	49.96	530	57.11	6	6233	50.45	628	67.67
H5N1	LCC	3178	1012	2884	90.75	229	22.63	2	2884	90.75	305	30.14
	CWC			2830	89.05	201	19.86	8	2878	90.56	327	32.31

We denote ItFS based on the algorithm LCC (*resp.*, CWC) by ItLCC (*resp.*, ItCWC). Then, Tables 1, 2, 3, 4, 5, 6, 7 and 8 claim the following statements.

1. For every segment, ItCWC can be always applied iteratively to the subtypes of

H1N1, H3N2 and H5N1 but not to the subtype of H2N2. On the other hand, ItLCC cannot be always applied iteratively to their subtypes.

2. For every segment and every subtype, the number of iterations in ItLCC is at most 2. However, $|e|$ does not change and just $|X|$ increases even if the number is 2.

3. For ItCWC, the number of iterations is more than 5 and less than 10 for the subtype of H1N1, more than 4 and less than 8 for the subtype of H3N2 and more than 2 and less than 10 for the subtype of H5N1. In particular, for the subtype of H5N1, the number of iterations in ItCWC is more than 5 except the segment of PA (that is 2).

4. For just the segment of PA, ItLCC and ItCWC can be always applied iteratively to all the subtypes of H1N1, H2N2, H3N2 and H5N1. However, almost number of iterations is 2 except ItCWC for the subtype of H1N1 (that is 7) and for the subtype of H3N2 (that is 8).

Concerned with Statement 1, we summarize the results of ItCWC for the 3 subtypes of H1N1, H3N2 and H5N1 as Table 9. Here, $\Delta|e|$ (*resp.*, $\Delta|X|$) denotes the difference of $|e|/m$ (*resp.*, $|X|/n$) between ItCWC and CWC.

Table 9: The results of ItCWC for the subtypes of H1N1, H3N2 and H5N1.

subt.	seg.	CWC				ItCWC					Δ	
		$\|e\|$	$\|e\|/m$	$\|X\|$	$\|X\|/n$	#	$\|e\|$	$\|e\|/m$	$\|X\|$	$\|X\|/n$	$\Delta\|e\|$	$\Delta\|X\|$
H1N1	PB2	8854	75.21	777	28.51	7	9011	76.55	1019	37.39	1.34	8.88
	PB1	8673	74.50	792	29.89	10	8865	76.15	1100	41.51	1.65	11.67
	PA	8803	74.72	807	21.14	7	8922	75.73	1012	26.51	1.01	5.37
	HA	23140	79.48	1004	42.58	9	23305	80.05	1359	57.63	0.57	15.05
	NP	7724	62.57	623	38.08	5	7828	63.41	789	48.23	0.84	10.15
	NA	17108	71.95	865	48.73	5	17196	72.32	1053	59.32	0.37	10.59
	MP	8437	48.47	562	45.95	8	8504	48.85	717	58.63	0.38	12.68
	NS	6971	54.96	570	57.46	6	7053	55.61	682	68.75	0.65	11.29
H3N2	PB2	8666	74.68	755	30.97	8	8863	76.38	1060	43.48	1.70	12.51
	PB1	8646	74.42	768	24.11	6	8822	75.93	1048	32.90	1.51	8.79
	PA	8391	72.51	790	29.70	8	8577	74.12	1066	40.08	1.61	10.38
	HA	22507	76.45	1007	44.70	7	22649	76.93	1308	58.06	0.48	13.36
	NP	7874	63.67	602	32.66	6	7972	64.47	749	40.64	0.80	7.98
	NA	12474	73.04	764	45.10	5	12596	73.75	942	55.61	0.71	10.51
	MP	8301	49.86	498	44.35	4	8347	50.14	597	53.16	0.28	8.81
	NS	6172	49.96	530	57.11	6	6233	50.45	628	67.67	0.49	10.56
H5N1	PB2	2762	96.64	189	7.79	5	2765	96.75	233	9.60	0.11	1.81
	PB1	2741	94.48	182	7.22	5	2750	94.79	231	9.17	0.31	1.95
	PA	2714	96.14	185	7.88	2	2716	96.21	201	8.56	0.07	0.68
	HA	5556	94.17	283	13.73	9	5588	94.71	439	21.30	0.54	7.57
	NP	2767	94.21	185	11.34	9	2785	94.82	287	17.60	0.61	6.26
	NA	4026	92.68	255	13.92	7	4052	93.28	387	21.12	0.60	7.20
	MP	2854	87.20	202	17.75	10	2905	88.76	355	31.20	1.56	13.45
	NS	2830	89.05	201	19.86	8	2878	90.56	327	32.31	1.51	12.45

Table 9 shows the following statements.

5. The average values of $\Delta|e|$ for H1N1, H3N2 and H5N1 are 0.85, 0.94 and 0.66, respectively. On the other hand, the average values of $\Delta|X|$ for H1N1, H3N2 and H5N1 are 10.70, 10.36 and 6.42, respectively.

6. Increasing $\Delta|e|$ is independent from increasing $\Delta|X|$ in general.

7. For the subtype of H3N2, the segment of PB2 has the maximum value of $\Delta|e|$ and $\Delta|X|$. For the subtype of H5N1, the segment of MP has the maximum value of $\Delta|e|$ and $\Delta|X|$. On the other hand, for the subtype of H1N1, the segment PB1 has the maximum value of $\Delta|e|$ but not $\Delta|X|$ and the segment HA has the maximum value of $\Delta|X|$ but not $\Delta|e|$. In particular, the segment PB1 has the third maximum value of $\Delta|X|$, but the segment HA has the sixth maximum value of $\Delta|e|$.

8. Whereas the segments of PB2, PB1 and PA have larger values of $\Delta|e|$ for the subtypes of H1N1 and H3N2, the segments of MP and NS have larger values of $\Delta|e|$ for the subtype of H5N1. In particular, they have larger values of $\Delta|X|$ for the subtype of H5N1.

5 Conclusion

In this paper, we have first formulated the consistency-based feature selection problem as combinatorial optimization problems. Next, based on the consistency-based feature selection algorithms of LCC [18] and CWC [19][20][21], we have designed the algorithm ITFS as an *iterative consistency-based feature selection*. Finally, we have applied ITFS to nucleotide sequences of influenza A viruses and evaluated the results. Hence, we have observed that the number of explanatory instances for 3 subtypes of H1N2, H3N2 and N5N1 and all the 8 RNA segments is always increasing by ITCWC, and that for all the 4 subtypes and the RNA segment PA is always increasing by both ITLCC and ITCWC.

One of the reason that the algorithm ITLCC has not achieved the purpose to avoid eliminating too many inconsistent instances stated in Statement 2 in Section 4 is that we fix a threshold δ to 0 and not find an appropriate value of δ which is a future work. Also, since the iterative consistency-based feature selection is based on the algorithms LCC [18] and CWC [19][20][21], it is a future work to apply an iterative method to the other algorithms.

The consistency-based feature selection problem as combinatorial optimization problems is mixed the minimization problem for the number of features to the maximization problem for the number of explanatory instances.
Then, it is a future work to analyze
the exact intractability of computing it, for example, Σ_p^2-hardness beyond NP-hardness and non-approximability. It is also a future work to investigate whether or not the dual problem is meaningful and, if so, then to design an efficient method.

Also, in this paper, we apply the algorithm ITFS to nucleotide sequences of influenza A viruses. Then, it is a future work to apply it to other data set and evaluate the results. Furthermore, as stated in the last of Section 4, the framework of iterative feature selection is possible to be useful to increase explanatory instances, so it is a future work to analyze it.

Recently, Shimamura and Hirata have extended the algorithms CWC and LCC by reselecting adjacent sets of feature sets [16] and by introducing the fluctuation into increasing order of symmetric uncertainty [17]. Then, it is a future work to incorporate these extensions with the iterative feature selection in this paper.

From the viewpoint of knowledge discovery, the selected features are possible to induce some rules for a data set of the form that pairs of features and their values imply a class label, so it is a future work to extract such rules from the algorithm ITFS.

Acknowledgments

This work is partially supported by Grant-in-Aid for Scientific Research 17H00762, 16H02870, 16H01743 and 15K12102 from the Ministry of Education, Culture, Sports, Science and Technology, Japan. The authors would thank to Prof. Kilho Shin at Gakushuin University to give insightful comments for the feature selection as the combinatorial optimization problem. Also they would thank to anonymous referees of Information Engineering Express to give valuable comments to revise this paper.

References

[1] E. Amaldi, V. Kann, "The complexity and approximability of finding maximum feasible subsystems of linear system," Theoretical Computer Science, vol. 147, 1995, pp. 181–210.

[2] E. Amaldi, V. Kann, "On the approximability of minimizing nonzero variables or unsatisfied relations in linear system," Theoretical Computer Science, vol. 209, 1998, pp. 237–260.

[3] A. Arauzo-Azofra, J. M. Benitez, J. L Castro, "Consistency measures for feature selection," Journal of Intelligent Information Systems, vol. 30, 2008, pp. 273–292.

[4] T. Bao, P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky, T. Tatusova, J. Ostell, D. Lipman, "The influenza virus resource at the National Center for Biotechnology Information," Journal of Virology, vol. 82, 2008, pp. 596–601. Also available at: `http://www.ncbi.nlm.gov/genomes/FLU/`.

[5] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanozos, "Feature selection for high-dimensional data," Springer, 2015.

[6] I. Guyon, S. Gunn, M. Nikravesh, L. Zadeh, "Feature extraction: Foundations and applications," Springer, 2006.

[7] I. Hamada, T. Shimada, D. Nakata, K. Hirata, T. Kuboyama, "Classifying nucleotide sequences and their positions of influenza A viruses through several kernels," Proc. 4th Int'l Conf. on Pattern Recognition Applications and Methods (ICPRAM 15), 2015, pp. 354–359.

[8] H. Liu, H. Motoda, M. Dash, "A monotonic measure for optimal feature selection," Proc. 10th European Conf. Machine Learning (ECML 98), 1998, pp. 101–106.

[9] S. Makino, T. Shimada, K. Hirata, K. Yonezawa, K. Ito, "A trim distance between positions in nucleotide sequences," Proc. 15th Int'l Conf. on Discovery Science (DS 12), Lecture Notes in Artificial Intelligence, vol. 7562, 2012, pp. 81–91.

[10] L. Molina, L. Belanche, A. Nebot, "Feature selection algorithms: A survey and experimental evaluation," Proc. 2002 IEEE Int'l Conf. Data Mining (ICDM 02), 2002, pp. 306–313.

[11] Z. Pawlik, "Rough set: Theoretical aspects of reasoning about data," Kluwer Academic Press, 1991.

[12] W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery: "Numerical recipes in C (2nd edition)," Cambridge University Press, 1992.

[13] C. E. Shannon: "A mathematical theory of communication," Bell System Technical Journal, vol 27, 1948, pp. 379–423, 623–666.

[14] T. Shimada, I. Hamada, K. Hirata, K. Yonezawa, K. Ito, "Clustering of positions in nucleotide sequences by trim distance," Proc. IIAI Int'l Conf. on Advanced Applied Informatics (IIAI AAI 13), 2013, pp. 129–134.

[15] S. Shimamura, K. Hirata, "On temporal and regional analysis for nucleotide sequences of influenza A (H1N1) viruses based on feature selection," Proc. 2016 Int'l Workshop on Smart Info-Media Systems in Asia (SISA 16), 2016, pp. 38–42.

[16] S. Shimamura, K. Hirata, "The reselection of adjacent sets by consistency-based feature selection," Proc. 2nd Int'l Conf. on Information Science and System (ICISS 19), 2019, 4 pages.

[17] S. Shimamura, K. Hirata, "Introducing fluctuation into increasing order of symmetric uncertainty for consistency-based feature selection," Proc. 15th Annual Conf. on Theory and Applications of Models of Computing, Lecture Notes in Computer Science, vol. 11436, 2019, pp. 550–565.

[18] K. Shin, X. M. Xu, "Consistency-based feature selection," Proc. 13th Int'l Conf. Knowledge-Based and Intelligent Information & Engineering (KES 09), 2009, pp. 342–350.

[19] K. Shin, D. Fernalndes, S. Miyazaki, "Consistency measures for feature selection: A formal definition, relative sensitivity comparison, and a fast algorithm," Proc. 20th Int'l Joint Conf. Artificial Intelligence (IJCAI 11), 2011, pp. 1491–1497.

[20] K. Shin, T. Kuboyama, T. Hashimoto, D. Shepard, "Super-CWC and super-LCC: Super fast feature selection algorithms," Proc. IEEE Int'l Conf. Big Data, 2015, pp. 61–67.

[21] K. Shin, S. Miyazaki, "A fast and accurate feature selection algorithm based on binary consistency measure," Computational Intelligence, vol. 32, 2016, pp. 645–667.

[22] Z. Zhao, H. Liu, "Searching for interacting features," Proc. 16th Int'l Joint Conf. Artificial Intelligence (IJCAI 07), 2007, pp. 1156–1161.