Online Streaming Feature Selection Based on Conditional Information Entropy

Huaming Wang , Guoyin Wang*, Xianhua Zeng and Siyuan Peng
Chongqing Key Laboratory of Computational Intelligence,
Chongqing University of Posts and Telecommunications, Chongqing 400065, P.R. China
Email: wanggy@cqupt.edu.cn

***Abstract*—Online streaming feature selection (OSFS) algorithms, producing an approximately optimal subset from so-far-seen features in real time, are capable of addressing feature selection issues in extreme large or even infinite dimensional space. There are several algorithms proposed carrying out in OSFS way. However, some of these algorithms need prior knowledge about the entire feature space which is inaccessible in real OSFS scenario. Besides, results of them are sensitive to the permutations of features. In this paper, we first propose an OSFS framework based on the uncertainty measures in rough sets theory. The framework needs no additional information, except for the given data. Moreover, a sorting mechanism is adopted in the framework, as creates its stability to varying the order of features. Then, specifying the uncertainty measure with conditional information entropy (CIE), we design an algorithm named CIE-OSFS based on the framework. Comprehensive experiments are conducted to verify the effectiveness of our method on several high dimensional benchmark data sets. The experimental results indicate that CIE-OSFS achieves more compactness with the prerequisite of guaranteeing the predictive accuracy and performs more stably to the changing of features' order than other algorithms in most cases.**

1. Introduction

Data collection has grown tremendously with the rise of Big Data applications [1]. Not only is the scale of instances growing larger, but also the number of features for an instance becomes enormous [2]. However, many of these features are irrelevant or redundant and which matters is not known. Unnecessary variables, causing the increment of calculation complexity and computation time, may greatly reduce the performance of classifiers. Therefore, it is essential to conduct feature selection procedure.

Traditional feature selection algorithms hold the hypothesis that the number of features is fixed and the entire feature space is accessible beforehand [3]. However, when it comes to medical learning problems, for example, emerging medical technologies bring new features, the entire feature space can be extreme large or even infinite. No prior knowledge about the entire space could be referred by the learning algorithms during the runtime. Moreover, these tasks demand high reliability of temporary results produced by online learning mechanism once new features is available. Wu et.al [4] formulated these feature selection tasks with dynamic features as online streaming feature selection (OSFS) problems. Aiming to deal with stream data, OSFS algorithms generate updated results in real time with previous results and new information but use no prior knowledge.

Although there are several methods [3], [4], [5], [6] carrying out in OSFS way, some of them used prior knowledge about the entire feature space. Besides, their results are related to the ordering of features. But in real streaming feature selection scenario, both the entire feature space and the related order of features are not accessible. Methods using prior knowledge or sensitive to the ordering of features may lead to their inefficiency in real applications. Hence, OSFS algorithms should not only use no prior knowledge to deliver credible results but robust to the changing of the order of features.

In this paper, we consider OSFS problems from the rough sets perspective and propose an uncertainty measures based framework to address OSFS issues. Rough sets [7] techniques, gaining insights into properties of given data, contribute its requiring no prior knowledge. In addition, a sorting mechanism enable the framework robust to the changing of the ordering of features. Specifying the uncertainty measure with conditional information entropy in rough sets theory, we design a OSFS algorithm named online streaming feature selection based on conditional information entropy (CIE-OSFS).

The contributions of this study include two aspects: (1) a stable online streaming feature selection framework based on uncertainty measures is designed; (2) an online streaming feature selection algorithm applying conditional entropy information (CIE) in rough sets theory [8] called CIE-OSFS is proposed.

2. Related Works

Feature selection, aiming to address "the curse of dimensionality" problem caused by high dimensionality data [9], is an active area of research in pattern recognition [10], machine learning [11] and data mining [12]. According to the relation between feature selection process and learning

978-1-5386-3120-1/17 $31.00 © 2017 IEEE
DOI 10.1109/ICBK.2017.44

algorithms, feature selection methods can be categorized into wrapper, embedded and filter methods.

The wrapper methods using the feedback of a specific classifier to determine which features to be selected. But their performance are limited to specific classifiers and the computation time of the wrapper methods is general high [4]. The embedded methods design classifiers attached with spares mechanisms to maximize classification accuracy and minimize the size of features used simultaneously. While the filter methods, use consistency or dependency to guide feature selection procedure, are independent to classifiers [3]. Compared to wrapper and embedded ones, the filter methods are more time efficient.

Most feature selection algorithms proposed, wrapper, embedded or filter ones, handle feature selection problems with the assumption that all candidate features are available beforehand [13]. But in OSFS tasks [14], the entire feature space may be increasing and could be extremely large or even infinite. Hence, traditional feature selection methods are unworkable in real streaming feature cases.

Several methods have been proposed to address streaming feature selection issues. Grafting [5] treats the feature selection task as a stream-wise regularized risk minimization problem. New features are selected if the improvement of accuracy made by them is greater than a predefined threshold. However, there is no mechanism to removing redundant features selected previously, rendering it suffering from the nesting effect. Alpha-investing [6] uses a linear regression to evaluate the modified model. It uses a p-value to determine new features which are selected or not. Furthermore, Alpha-investing and Grafting used prior information about the structure of feature space [4], which is impossible to obtain on the original streaming tasks. Hence they might not produce good performance in real applications.

Wu [4] proposed the fast-OSFS algorithm, needing no prior knowledge about the feature space, which contains two major steps: online relevance analysis and online redundancy analysis. The first step discards irrelevant features and the second eliminates redundant features. However, its examination order of redundant features is related to the relative order of features.

In real streaming feature selection scenario, both the feature space and the related order of features are not accessible. Methods using prior knowledge or sensitive to the ordering of features may lead to their inefficiency in real applications. In this paper, we propose an OSFS framework suitable for real OSFS applications, which is robust to the ordering of features and requires no prior knowledge except for given data itself.

3. An OSFS Framework Based on Rough Sets Using Uncertainty Measures

Rough set theory is an effective soft computation technique [15] [16] for applications with vagueness and uncertainty. It has been applied very fruitfully in many real-life applications to reduce the dimensionality. Best of all, rough sets based feature selection techniques deliver the dependencies and significance of variables in the condition of gaining insights into properties of given data. It uses no additional information or prior knowledge about the feature space, as is suitable for real streaming feature scenarios.

3.1. Traditional Feature Selection Methods Using Uncertainty Measures

An information system can be formulated as $IS = (U, A, V, f)$, where $U = \{x_1, x_2, ..., x_n\}$ is a nonempty and finite set of samples. $A = C \cup D$ is a finite set of attributes, where $C = \{a_1, a_2, ..., a_n\}$ is the set of conditional attributes (features) and $D = \{d\}$ is the set of decision attributes (label). In addition, $V = \bigcup_{a \epsilon A} V_a$, where V_a is the domain of a. $f : U \rightarrow V_a$ (for every $a \epsilon A$) is the information function.

Definition 1 (Uncertainty Measure) For any set $B \subseteq A$, the indiscernible relation [8] defined on B can be formulated as $IND(B) = \{(x, y) \epsilon U \times U | \forall a \epsilon B, f(x) = f(y))\}$. Accordingly, the partitions of U on B is denoted by $U/B = \{X_1, X_2, ..., X_m\}$. The uncertainty measure of U on B is defined as $UM(B)$ which measures the discernibility or the information contributed by B. In addition, the uncertainty measure function B can be specified as different entropies and granularities [17] in rough sets. For example, when the information entropy [8] is used, the uncertainty of $\boldsymbol{U}$ on $\boldsymbol{B}$ is

$$UM(B) = H(B) = -\sum_{i=1}^{n} p(X_i) log(p(X_i)) \quad (1)$$

where $X_i \epsilon U/B$ and $p(X_i) = |X_i|/|U| (i = 1, 2, ..., m)$.

Definition 2 (Conditional and Joint Uncertainty) Let $\boldsymbol{P}$ and $\boldsymbol{Q}$ be the set of conditional attributes and $P, Q \subseteq A$. The joint uncertainty of $\boldsymbol{P}$ and $\boldsymbol{Q}$ is defined as

$$JUM(P, Q) = UM(P \cup Q) \quad (2)$$

And the conditional uncertainty of $\boldsymbol{Q}$ on $\boldsymbol{P}$ is

$$CUM(Q|P) = JUM(P, Q) - UM(P) \quad (3)$$

$JUM(P, Q)$ measures the discernibility of $\boldsymbol{P}$ and $\boldsymbol{Q}$, while $CUM(Q|P)$ represents the information that $\boldsymbol{Q}$ carries on the condition of $\boldsymbol{P}$.

Definition 3 (Reducible attribute or Redundant feature) for any attribute $s \in A - P$, the significance of $\boldsymbol{s}$ on the condition of $\boldsymbol{P}$ is

$$sig(s, P, d) = UM(\{d\}|P \cup \{s\}) - UM(\{d\}|P) \quad (4)$$

When its value equals to 0, $\boldsymbol{s}$ contributes no discernible information to $\boldsymbol{P}$. Hence, $\boldsymbol{s}$ is redundant in $\boldsymbol{P}$, if and only if

$$sig(r, P, d) = 0 \quad (5)$$

Definition 4 (Reduct) $\boldsymbol{P}$ is a reduct of $\boldsymbol{A}$, if and only if

$$UM(\{d\}|P) = UM(\{d\}|A) \quad (6)$$

where $d \epsilon A$ is the label. Besides, for any $r \in P$,

$$sig(r, P, d) \neq 0 \quad (7)$$

Traditional feature selection methods using uncertainty measures [8], [18], [19] generally have four steps:

1) initialize $SF = \{\}$ and $C = \{a_1, a_2, ..., a_n\}$;
2) find $a \epsilon C$ by maximizing a specific uncertainty measure;
3) $SF = SF \cup a$, $C = C - a$, repeat step 2 until stop criteria met;
4) return selected features (reduct) SF.

Obviously, the complexity of its generating a reduct is $O(m^2)$, where $m = |C|$.

3.2. An OSFS Framework Using Uncertainty Measures

Simply, streaming feature selection issues can be addressed by maintaining a subset of selected features and updating it dynamically with new features flowing in. However, it is very time consuming in real applications for the existence of many irrelevant features [20] as Figure 1 shows. If irrelevant features flow in, rebuilding a reduct is very time consuming and unnecessary. Hence, we use the independence test filtering irrelevant features to improve time efficiency. The OSFS framework using the uncertainty measures is designed in Figure 2.

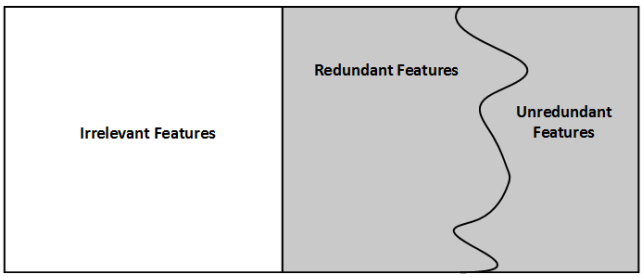

Figure 1. The feature space of real applications

The framework considers the streaming feature selection problems from rough sets perspective. It aims to dislodge redundant features while keeps the discernibility other than the classification accuracy or conditional probability fixed. There are two major steps in the framework. In the first step, the independence test is adopted to determine whether features are related to the target and irrelevant features are discarded. The next step applies uncertainty measures in rough sets theory to remove redundant features once new relevant features are added. Until there have no new features or the stopping criteria is met, the two steps are repeated. Moreover, the framework adopts a sorting mechanism to eliminate effects caused by the related order of features, as contributes its stability to the changing of features' order.

3.3. On Removing Redundant Features In The Framework

When a new feature related to decision attribute arrives, the framework sorts features by correlation coefficients before generating a reduct of current selected features. The sequence produced by sorting is used to guide attribute reduction process heuristically. During checking the redundancy

1. Initialization
   a) Selected features set $\boldsymbol{SF = \{\}}$, decision attribute (or label) $\boldsymbol{D}$
2. Discarding Irrelevant features
   a) Get a new feature $\boldsymbol{f}$
   b) Carrying out the independence test of $\boldsymbol{f}$ and $\boldsymbol{D}$
      - If $\boldsymbol{f}$ is an irrelevant feature. Discard $\boldsymbol{f}$
      - If not, $\boldsymbol{SF = SF \cup f}$
3. Removing Reducible (Redundant) features
   a) Sorting features of $\boldsymbol{SF}$ in descending order with correlation coefficient.
   b) $\boldsymbol{RED = \{\}}$, Add irredundant features to $\boldsymbol{RED}$ using uncertainty measures by Definition 3 in the order generated in a).
   c) Replace $\boldsymbol{SF}$ with its reduct $\boldsymbol{RED}$ generated by step b)
4. Repeat steps 2 and 3 until no new features or the stopping criteria is met
5. Output the selected features set $\boldsymbol{SF}$

Figure 2. An OSFS framework based on rough sets

of features, attributes with greater correlation coefficients have higher priority. No matter what the ordering that the features arrived is, the algorithm holds the most informative ones. Moreover, attributes with high correlation coefficients as well but contributing nothing to candidate selected attributes are discarded. While attributes with low correlation coefficient but contributing distinctive information are selected. Therefore, the framework is stable to the changing of features' order. In addition, every feature is checked only once in the order of sorting and irrelevant features are removed before reduction, as makes the framework much more efficiency than traditional feature selection methods using uncertainty measures illustrated in Section 3.1.

4. An OSFS Algorithm Using Conditional Information Entropy

Attribute reduction algorithms, using conditional information entropy (CIE) or other uncertainty measures, objective to hold discernibility with selected features. Compared to traditional attribute reduction algorithms intend to keep the positive region or the lower approximation unchanged [21], they have better fault tolerability [22]. In this section, we adopt conditional information entropy to measure the uncertainty and propose an online streaming feature selection algorithm under the framework illustrated in Section 3.2. Besides, a brief analysis of the efficiency of the proposed algorithm is given.

4.1. An Online Streaming Feature Selection Algorithm

The pseudo-code of the online streaming feature selection based on conditional information entropy (CIE-OSFS

in short) is shown in Algorithm 1.

Algorithm 1: The CIE-OSFS Algorithm

Input: decision attribute $\boldsymbol{d}$, features stream $\boldsymbol{F} = \{\boldsymbol{f_1}, \boldsymbol{f_2}, ..., \boldsymbol{f_n}\}$
Output: set of selected features $\boldsymbol{SF}$

$\boldsymbol{SF} = \{\}$;
repeat
  get $\boldsymbol{f_i}$ from the stream $\boldsymbol{F}$;
  carry out the independence test $Ind(f_i, d)$;
  if $\sim Ind(f_i, d)$ **then**
    $SF = SF \cup \{f_i\}$;
    /*using CIE to remove redundant features*/
    $RED = \{\}$; /*the set of selected features*/
    /*sort attributes by correlation coefficient */
    get $B = \{b_1, b_2, ..., b_m\}$; $\forall b_i, b_j \in SF$, $i > j$, s.t. $corr(b_i, d) \geq corr(b_j, d)$;
    calculate $CIE(\{d\}|SF)$;
    /* generating a reduct of SF */
    for $i = 1, 2, ...m$ **do**
      /* removing redundant features */

$$sig(b_i, RED, d) = CIE(\{d\}|RED) - CIE(\{d\}|RED \cup \{b_i\})$$

      if $sig(b_i, RED, d) > 0$ **then**
        $RED = RED \cup \{b_i\}$;
        if $CIE(\{d\}|RED) == CIE(\{d\}|SF)$ **then**
          break;
        end
      end
    end
    $SF = RED$;
  end
until *no new features or the accuracy threshold met*;

4.2. The Time Efficiency of CIE-OSFS

Suppose $\boldsymbol{R_t}$ is the selected feature subset at time $\boldsymbol{t}$ and $\boldsymbol{r} = |\boldsymbol{R_t}|$ is the number of features that $\boldsymbol{R_t}$ contains. Let $\boldsymbol{F_t}$ be the features arrived before time $\boldsymbol{t}$. $\boldsymbol{F_t^r}$ is a subset of $\boldsymbol{F_t}$ containing all features which are not independent to the decision attribute. Let $\boldsymbol{m_1} = |\boldsymbol{F_t}|$ be the number of features in $\boldsymbol{F_t}$ and $\boldsymbol{m_2} = |\boldsymbol{F_t^r}|$. The time complexity of CIE-OSFS consists of two parts: the complexity of conducting relevance analysis and the complexity of removing redundant features. Therefore, the average time complexity of CIE-OSFS is $O(m_1 + m_2 \cdot r^2)$. When all features are discarded in the independence test, the best time complexity is $O(m_1)$. While all features passed the independence test, the worst-case complexity is $O(m_1 \cdot r^2)$.

5. Experimental Results

In this section, we provide several experimental results to compare the effectiveness and stability of CIE-OSFS with fast-OSFS [4], Alpha-investing [6] and Grafting [5]. Comparative algorithms compared were performed using their original implementations and settings. The details of high dimensional benchmark data sets used are shown in Table 1. All data sets are from medical or healthy field. The predictive accuracy was produced by KNN classifier (k=3) using selected features. More results using other classification methods and comparisons of time efficiency can be found on Github[1]. All experimental results are obtained on a personal computer with Windows 10, Inter(R) Core (TM) i3-4170 CPU (3.70 GHz) and 8.00 GB memory employing MATLAB R2015a platform.

TABLE 1. DETAILS OF MEDICAL DATA SETS

No.	Dataset	Instances	Features	Source
1	prostate	102	12600	mldataorg
2	central-nervous-sys	60	7129	mldataorg
3	lung-cancer-michigan	96	7129	mldataorg
4	leu	38	7129	libsvm
5	marti0	500	1024	ChaLearm
6	reged0	500	999	ChaLearm
7	arcene	100	10000	NIPS2003
8	madelon	2000	500	NIPS2003

5.1. Comparisons of Results in OSFS Scenario

To simulate the scenario of online streaming feature selection and evaluate the effectiveness OSFS algorithm, we adopt prediction accuracy and the size of selected features to evaluate the process of streaming feature selection. In the experiments, the first 10 percentage of features were handled at beginning and the rest of features were processed increasingly. Moreover, 10-fold cross validation is employed. Figure 3 and Figure 4 give the performance of four algorithms on 8 data sets.

In Figure 3, CIE-OSFS obtains best performance among four methods on leu and madelon. Besides, CIE-OSFS outperforms Grafting on lung-cancer-michigan and performs better than fast-OSFS and Alpha-investing on arcene. On prostate, marti0 and reged0, neither of our method nor other three algorithms express definite advantage. As Figure 4 shows, CIE-OSFS selects less features other algorithms on all data sets except for madelon Moreover, CIE-OSFS selects least size of features with higher stability while features flow in.

Based on the observations above, we could conclude that CIE-OSFS delivers better compactness with competitive accuracy than other state-of-the-art algorithms when the entire feature space is not accessible or exhaustively searching over the entire feature space is impossible.

5.2. Comparisons of Stability

To test the stability of OSFS algorithms, we generate the meta-test ten times with the features in different random orders. In each meta-test, the data is divided into training

1. https://github.com/Hua-ming/CIE-OSFS

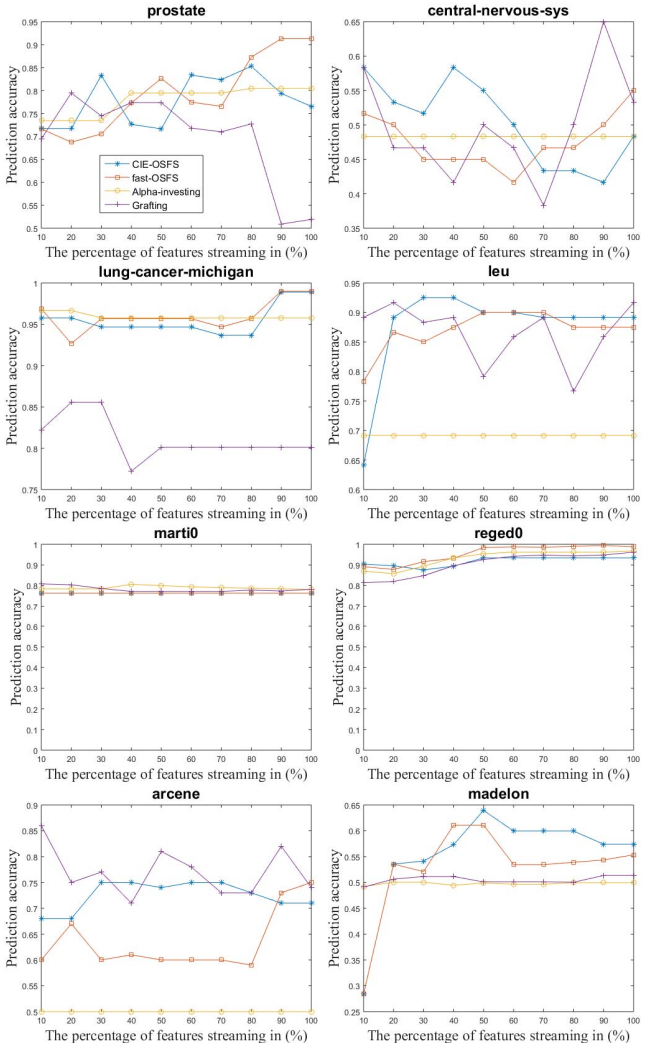

Figure 3. The prediction accuracy changes with respect to the number of features streaming in

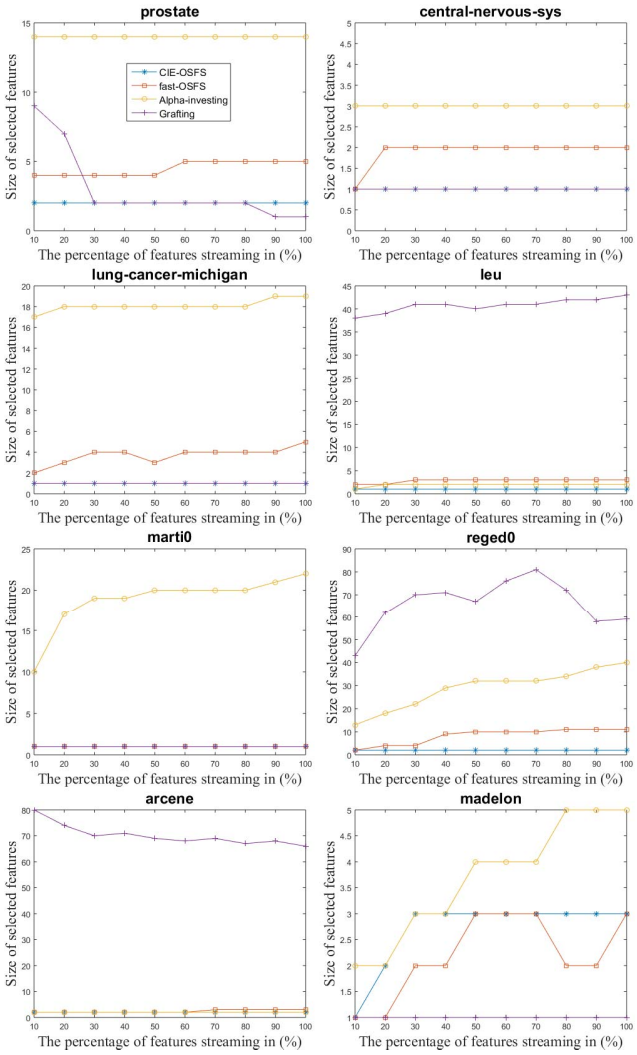

Figure 4. The size of selected features changes with respect to the number of features streaming in

(90%) and testing subset (10%) randomly. The training set is used to carry out feature selection tasks and train a classifier. Then, the testing set is given to the classifier to produce prediction accuracy. Table 2 gives the range (marked as r.) and the mean (marked as m.) of the size of selected features and Figure 5 shows the mean accuracy of ten trails on 8 data sets.

Figure 5 shows that CIE-OSFS procures better prediction accuracy than Alpha-investing and Grafting on all data sets. Compared to fast-OSFS, CIE-OSFS achieves better accuracy on 5 out of 8 data sets. In Table 2, CIE-OSFS outperforms Alpha-investing on all data sets and selects less features than fast-OSFS except for the marti0. Moreover, CIE-OSFS outperforms Grafting on 4 out of 8 data sets and delivers same results on other 4 data sets. Above all, the size of selected features delivered by CIE-OSFS varies minimum, while results of other algorithms are with wider

TABLE 2. COMPARISON OF SELECTED SUBSET STABILITY WHILE THE ORDER OF FEATURES CHANGING

No.	CIE-OSFS		fast-OSFS		Alpha-investing		Grafting	
	r.	m.	r.	m.	r.	m.	r.	m.
1	**0**	**2**	2	5.4	6	10.1	**0**	**2**
2	1	1.2	2	3	4	2.6	**0**	**1**
3	**0**	**1**	2	3.9	13	21.5	**0**	**1**
4	**0**	**1**	2	3.8	4	2.3	8	43.2
5	**0**	**1**	**0**	**1**	35	24.2	**0**	**1**
6	**0**	**2**	2	11.2	13	36.3	37	51.3
7	**1**	**2.4**	**1**	4.3	4	2.1	10	67.6
8	**0**	**3**	3	3.6	3	5.4	335	34.7

fluctuations. For example, on madelon, the size of selected features using Grafting varies with the range=335 in ten trails, but CIE-OSFS generates same size of features in all trails. These comparisons illustrate that CIE-OSFS generated more stable results than other three algorithms.

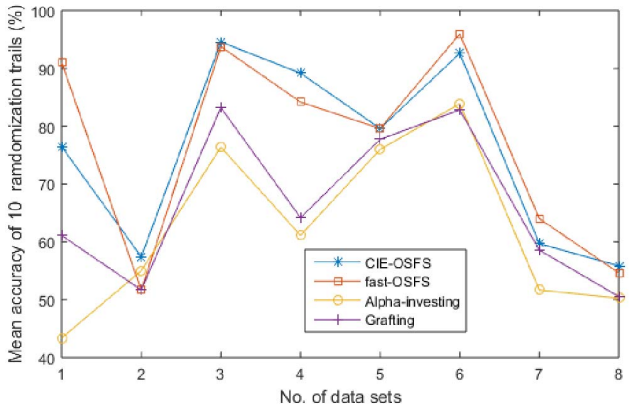

Figure 5. Comparison of accuracy stability while the order of features changing

6. Conclusion

In this paper, we propose an framework based on the uncertainty measures to address online streaming feature selection issues. The framework considers streaming feature selection problems in the view of rough sets, using the uncertainty measures and a sorting mechanism to select features heuristically. The proposed framework is stable to changing of features' order and requires no prior knowledge on the entire feature space. Then, a stable online streaming feature selection algorithm called CIE-OSFS was designed based on this framework.

Comparisons have shown that CIE-OSFS demonstrates more compactness with the prerequisite of guaranteeing the predictive accuracy than other three state-of-the-art algorithms, fast-OSFS, Alpha-investing and Grafting. Moreover, CIE-OSFS delivers more stable results than other algorithms under the changing of features' order. Number of instance may increase as features flow in, we will study how to perform online streaming feature selection under the condition that the number of samples increase in our future work.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (Grant no. 2016YFB1000905), the National Natural Science Foundation of China (Grant nos. 61672120, 61572091), the Chongqing Natural Science Foundation Program (Grant no. cstc2015jcyjA40036) and the Chongqing Postgraduate Scientific Research and Innovation Projects of China No.CYS16170.

References

[1] X. D. Wu, H. H. Chen, G. Q. Wu, and J. Liu, "Knowledge engineering with big data," *IEEE Intelligent Systems*, vol. 30, no. 5, pp. 1–1, 2015.

[2] M. Bennasar, Y. Hicks, and R. Setchi, "Feature selection using joint mutual information maximisation," *Expert Systems with Applications An International Journal*, vol. 42, no. 22, pp. 8520–8532, 2015.

[3] S. Eskandari and M. M. Javidi, "Online streaming feature selection using rough sets," *International Journal of Approximate Reasoning*, vol. 69, no. C, pp. 35–57, 2015.

[4] X. D. Wu, K. Yu, W. Ding, H. Wang, and X. Zhu, "Online feature selection with streaming features." *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 35, no. 5, pp. 1178–1192, 2013.

[5] S. Perkins and J. Theiler, "Online feature selection using grafting," in *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*, 2003, pp. 592–599.

[6] J. Zhou, D. Foster, R. Stine, and L. Ungar, "Streaming feature selection using alpha-investing," in *Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, Usa, August*, 2005, pp. 384–393.

[7] Z. Pawlak, J. Grzymala-Busse, R. Slowinski, and W. Ziarko, "Rough sets," *International Journal of Parallel Programming*, vol. 11, no. 5, pp. 341–356, 1982.

[8] G. Y. Wang, H. Yu, and D. C. Yang, "Decision table reduction based on conditional information entropy," *Chinese Journal of Computers*, vol. 25, no. 7, pp. 759–766, 2002.

[9] H. W. Liu, J. G. Sun, L. Liu, and H. J. Zhang, "Feature selection with dynamic mutual information," *Pattern Recognition*, vol. 42, no. 7, pp. 1330–1339, 2009.

[10] X. Zhang, C. L. Mei, D. G. Chen, and J. H. Li, "Feature selection in mixed data: A method using a novel fuzzy rough set-based information entropy," *Pattern Recognition*, vol. 56, no. 1, pp. 1–15, 2016.

[11] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1, pp. 389–422, 2002.

[12] Z. Zhao and H. Liu, "Semi-supervised feature selection via spectral analysis," in *Siam International Conference on Data Mining, April 26-28, 2007, Minneapolis, Minnesota, Usa*, 2007.

[13] M. ShakilPervez and D. Md. Farid, "Literature Review of Feature Selection for Mining Tasks," *International Journal of Computer Applications*, vol. 116, no. 21, pp. 30–33, Apr. 2015.

[14] J. Wang, Z. Q. Zhao, X. G. Hu, Y. M. Cheung, M. Wang, and X. D. Wu, "Online group feature selection," in *International Joint Conference on Artificial Intelligence*, 2014, pp. 1757–1763.

[15] J. W. Guan and D. A. Bell, "Rough computational methods for information," *Artificial Intelligence*, vol. 105, no. 1-2, pp. 77–103, 1998.

[16] G. Y. Wang, *Rough set theory and knowledge acquisition*. Xi'an Jiaotong University Press, 2001.

[17] J. Y. Liang, Z. Shi, D. Li, and M. J. Wierman, "Information entropy, rough entropy and knowledge granulation in incomplete information systems," *International Journal of General Systems*, vol. 35, no. 6, pp. 641–654, 2006.

[18] G. Y. Wang, "Rough reduction in algebra view and information view," *International Journal of Intelligent Systems*, vol. 18, no. 6, pp. 679–688, 2003.

[19] F. Wang, J. Y. Liang, and Y. H. Qian, "Attribute reduction: A dimension incremental strategy," *Knowledge-Based Systems*, vol. 39, no. 2, p. 95108, 2013.

[20] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," *Machine Learning Proceedings*, pp. 121–129, 1998.

[21] G. Y. Wang, J. Zhao, J. J. An, and Y. Wu, "A comparative study of algebra viewpoint and information viewpoint in attribute reduction." *Fundamenta Informaticae*, vol. 68, no. 3, pp. 289–301, 2005.

[22] Q. H. Zhang, J. J. Yang, and L. Y. Yao, "Attribute reduction based on rough approximation set in algebra and information views," *IEEE Access*, vol. 4, pp. 5399–5407, 2016.