International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.8, No.11 (2015), pp.433-444
http://dx.doi.org/10.14257/ijsip.2015.8.11.39

A Feature Selection Algorithm based on Hoeffding Inequality and Mutual Information

Chunyong Yin[1], Lu Feng[1], Luyu Ma[1], Zhichao Yin[2] and Jin Wang[1]

[1]*School of Computer and Software, Jiangsu Key Laboratory of Meteorological Observation and Information Processing, Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science & Technology, Nanjing 210044, China*
[2]*Nanjing No.1 Middle School, Nanjing 210001, China*

Abstract

With the rapid development of the Internet, the application of data mining in the Internet is becoming more and more extensive. However, the data source's complex feature redundancy leads that data mining process becomes very inefficient and complex. So feature selection research is essential to make data mining more efficient and simple. In this paper, we propose a new way to measure the correlation degree of internal features of dataset which is a mutation of mutual information. Additionally we also introduce Hoeffding inequality as constraint of constructing algorithm. During the experiments, we use C4.5 classification algorithm as test algorithm and compare HSF with BIF(feature selection algorithm based on mutual information). Experiments results show that HSF performances better than BIF[1] in TP and FP rate, what's more the feature subset obtained by HSF can significantly improve the TP, FP and memory usage of C4.5 classification algorithm.

Keywords: *Hoeffding inequality, data-stream, feature selection, mutual Information*

1. Introduction

1.1. Research Background

Since the beginning of twenty-first Century, data processing object related machine learning and data mining research shows a high latitude and large amount of data characteristics. (while mining the user data in the electronic commerce website, each user's product information often achieves the thousands of kinds and these data are sparse and dispersed, such a dimension and the degree of sparsity can hardly be found in the user data which greatly reduces the utilization of data mining; in the field of intrusion detection, the higher dimension of data, detection algorithms required for the processing of time is longer, memory on the computer is bigger, resulting in detection results are not accurate or not in time.) This is called "dimension disaster". On the other way, too many irrelevant features will result in the redundancy of data storage and reduce the efficiency of computer processing. Therefore, the improvement of the efficiency of data mining and machine learning algorithm is great in the data mining and machine learning algorithm research.

The ultimate goal of feature selection is to obtain the subset of the initial feature set, the dimension of the subset is as low as possible and the feature of the subset is also able to describe the data. The first step is to select the optimal feature subset selection criteria appropriate data, and then according to the data evaluation criteria are specific to the specific classification algorithm can be divided into three models of feature selection algorithm: Filter Model, Wrapper Model and Embeded Model [2]. The main research of

ISSN: 2005-4254 IJSIP
Copyright ⓒ 2015 SERSC

this paper is based on the Filter model. The Filter model is based on the structure of the data within the structure of the characteristics of the most relevant characteristics.

The typical algorithm of Filter model is based on the distance criterion RELIEF[3]. According to the correlation between each feature and the category of RELIEF, the weight of feature is lower than the threshold, but the RELIEF algorithm is only suitable for the two kinds of data sets, the limitations are relatively large. The RELIEF is extended to the RELIE-F[4, 5] algorithm based on Kononeill. But Sun[6, 7] prove the weights of each feature RELIE-F algorithm to get the easy to be disturbed by the data noise and RELIE-F nearest neighbor selection where space is not consistent with weighted vector space.

The feature selection method based on correlation measurement is mainly to evaluate the correlation between the features of each other [8-13]. The maximum correlation criterion is Pearson statistics, entropy, symmetric uncertainty criterion and mutual information criterionin the feature selection method based on correlation metric. The Pearson statistic is a linear dependence relationship between features, entropy, symmetric uncertainty and mutual information criterion are derived from information theory which can be used to measure the nonlinear relationship and linear relationship. Hall uses the Pearson statistics and the symmetry of the uncertainty criterion to evaluate the correlation degree between the feature and the sample type, and get the feature selection algorithm based on the correlation degree CFS[14]; Liu proposed a fast filtering algorithm based on correlation degree FCBF[15], FCBF uses the symmetric criterion to select the relevant features, and then uses the Balnket Markov concept to filter the redundant features in the relevant features; Guo and Nixon[16] use mutual information criterion to evaluate the characteristics of relevance and redundancy degree. It is proved that the proof of joint mutual information can be simplified for pairwise feature between the expressions of mutual information and reduce the computational complexity.

General feature selection is based on static data sets, the focus of the study is also inclined to data flow with the development of the Internet in recent years.

Different from the traditional static data block model, the data stream has the following characteristics:

1) High speed, real-time;

2) Large data size, not to be stored in memory or hard disk;

3) Data is reproduced, the cost of storage data is expensive, only a single scan data, unless specifically required, otherwise it will not store data;

4) The information contained in the data will be changed at any time.

In order to solve the adaptability of the algorithm to the data flow, the following conditions should be satisfied first:

1) The space complexity of the algorithm must be independent of the sample number of the data stream;

2) The algorithm must be able to adapt to the changing of the data stream.

In order to meet the above requirements, this paper adopts the idea of Hoeffding [17, 18] inequality to construct the algorithm which avoids the spatial complexity of the algorithm, because the Hoeffding inequality is based on the statistical basis which can be adapted to the changing of the data stream at the same time. To measure the degree of correlation between different set of inner elements, this paper proposes a new way based on mutual information which can take into account the correlation between features.

 Copyright ⓒ 2015 SERSC

1.2. Paper Work

HSF algorithm is proposed in this paper. HSF algorithm improve the mutual information measure to measure the correlation between the characteristics of the collection based on mutual information and the suitable feature subset is selected based on the statistical Hoeffding inequality. HSF further reduces data transmission, storage and processing redundancy and memory usage. Besides that HSF also reduces the time of data processing.

2. Dynamic Data-Stream Feature Selection Algorithm

2.1. Set Unit Mutual Information

Mutual information is a useful measurement in the information theory. It can be viewed as a random variable that contains information about another random variable, or a random variable that is reduced by a given random variable. Generally speaking, there is always noise and interference in the channel, the information sources emit message x, the sink can only receive the message as a result of the interference caused by the action of a deformation of the through the channel. The sink surmises the probability of receiving that the source sink issued, this process can be described by a posteriori $P(x \mid y)$. Accordingly, the probability of x, $P(x)$ is called a priori probability.

We define the mutual information of to x as below:

$$I(x, y) = \sum_{x,y} P(x, y) \log \frac{P(x, y)}{P(x) P(y)}$$

On the basis of this, the paper puts forward the set unit mutual information which measures the mutual influence between the two features in the set. Assuming set $A = \{x_1, x_2, \ldots x_n\}$, there are C_n^2 ways of arranging the elements in the set A. Then we define set unit mutual information as below:

$$I(A) = \left| \frac{\sum_{j=1}^{n} \sum_{i=2}^{n} I(x_j, x_i)}{C_n^2 - n} \right|$$

Set unit mutual information reflects the correlation degree between the two feature elements in the set, different sets unit mutual information shows the difference of the correlation between the elements in the set. For the collection A and B, if $I(A) > I(B)$, it is usually described that the collection of A is more suitable for reflecting the real information of a data than the feature elements contained in the B.

2.2. Hoeffding Inequality Theory

Hoeffding bound depicts the true possibility of the event A, n independent repetition frequency differences observed in Bernouli experiment and application in many different experiments. Hoeffding inequalities are given a probability boundary, for the data sample, we can get the real description of the event in a small range error under the confidence $1 - \delta$.

If $\bar{F}(x) - \bar{F}(y) > \sqrt{\frac{R^2 \ln(1/\delta)}{2m}}$, ($R$ is range of x and y, m is data samples), then the event description error of x is less than y. For feature selection, it is very important that how to select as little as possible to describe the real situation of the data. In order to

control the error of the continuous variation of the data stream, we must ensure that the error probability of the selection process is smaller in the real data stream environment.

2.3. HSF Algorithm Description

Under the given data stream environment, each data format is $X=\{x_1,x_2,...x_n\}$, we take unit set mutual information as a measure of the intrinsic feature association, all the subsets of X are listed, and the average values of the set unit mutual information of different sets are calculated. If the Hoeffding inequality is satisfied, it can be concluded that the collection is more accurate under the confidence $1-\delta$.

HSF algorithm steps:

Input:

Data sample $X=\{x_1,x_2,...x_n\}$;

Confidence $1-\delta$;

Set unit mutual information formula $I(A)=|\frac{\sum_{j=1}^{n}\sum_{i=2}^{n} I(x_j,x_i)}{C_n^2-n}|$;

Output:

Subset of X.

1. Listing all the subsets of X (except single element) $X_1,X_2,...X_k$, goto2;

2. Initializing the set unit mutual information $I(X_1),I(X_2),...I(X_k)$ of $X_1,X_2,...X_k$ and data samples as 0, goto3;

3. When data arrival $m+1$, calculating the mean value of $\bar{I}(X_1),\bar{I}(X_2),...\bar{I}(X_k)$, $\bar{I}(X_k)=\frac{I(X_k)}{m}$, goto4;

4. Choosing the biggest 2 of $\bar{I}(X_1),\bar{I}(X_2),...\bar{I}(X_k)$, $\bar{I}(X_a)$ and $\bar{I}(X_b)$, if $\bar{I}(X_a)-\bar{I}(X_b)>\sqrt{\frac{R^2\ln(1/\delta)}{2m}}$ goto5; else goto3;

5. Outputting X_a.

2.4. Algorithm Analysis

In the face of data stream environment, HSF algorithm can adapt to the characteristics of the data stream continuously, and the calculation of the logarithmic data can only be used to scan the data. In the selection of feature subsets, the transmission efficiency of the data stream is not impacted. The space complexity of HSF algorithm is $O(n^2)$, is the total number of features in the data sample. Therefore, the HSF algorithm's memory usage is not increasing with the data sample size increasing which greatly adapt to the dynamic data-stream.

 Copyright ⓒ 2015 SERSC

3. Experiments and Results Analysis

3.1. Experiment Dataset

KDD CUP99 intrusion detection dataset is the classic dataset to test the intrusion detection system comprehensively and also the most influential and credibility dataset in the academia presently. The 10% part of the KDD CUP99 is the most commonly used in the network security intrusion detection research.

KDD CUP99 dataset include 9 basic network connection features, 13 network connection content features, 19 network traffic features. The data type of these features has two kinds of continuous and discrete.

3.2. Experiment Scheme and Evaluation Standards

Experiment-1: we bring the 100, 500, 1000, 5000, 10000, 50000, 100000 samples situation to the experiment as the experiment data-source. HSF and BIF are used to get subset of the features after processing the data-source. Then we use C4.5 classification algorithm to classify the data-source only contains the features subset. The experiment mainly focuses on comparison of the time, TP rate, FP rate and the memory usage between HSF and BIF.

Experiment-2: we bring the 100, 500, 1000, 5000, 10000, 50000, 100000 samples situation to the experiment as the experiment data-source. HSF is used to get subset of the features after processing the data-source. Then we use C4.5 classification algorithm to classify the data-source only contains the features subset. The experiment mainly focuses on comparison of the time, TP rate, FP rate and the memory usage between original data-source and processed data-source.

3.3. Experiment Result Analysis

Experiment-1: Figure 1 shows that after processed by the C4.5 classification algorithm, the classification accuracy is improved with the increase of sample number using both data-source contains the BIF and HSF selected features subset. The magnitude begins to be flat at the 500 point. We can confirm that the subset of HSF is better than the subset of BIF by comprising the 2 classification accuracy classified by C4.5 algorithm. Because the HSF does not only consider the feature attributes for the entire data, but also considers the correlation between each feature. So the subset of HSF can reflect the real situation of the data better than BIF.

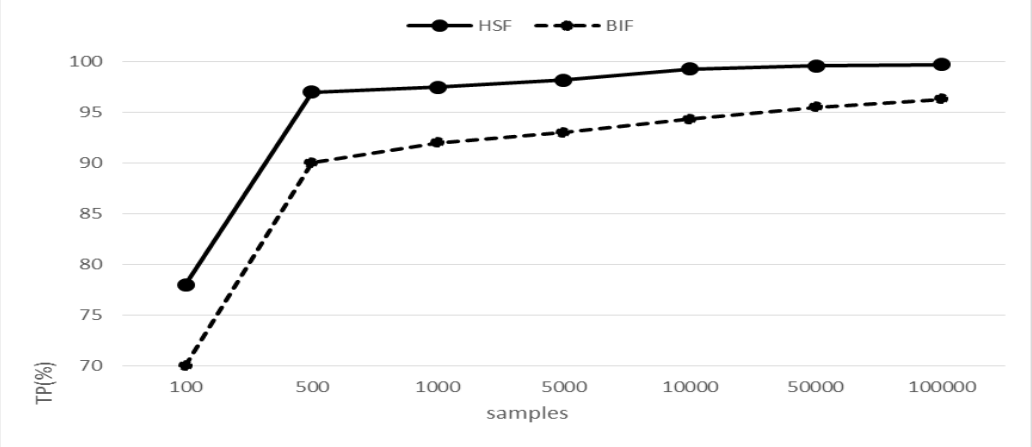


Figure 1. HSF and BIF Feature Selection Algorithm to Deal with the Same Amount of Data after the C4.5 Algorithm Classification Accuracy Comparison

Copyright ⓒ 2015 SERSC

Figure 2 shows that after processed by the C4.5 classification algorithm, the classification accuracy is reduced rapidly with the increase of sample number using both data-source contains the BIF and HSF selected features subset. However, with the increase of the data samples, the classification error rate tends to slow down, and the change is not obvious. The results of the same number of samples HSF algorithm is better than BIF, but with the increase of sample size, the gap between the two is also reduced tightly. Because the HSF does not only consider the feature attributes for the entire data, but also considers the correlation between each feature. So the subset of HSF can reflect the real situation of the data better than BIF and the experiment FP of HSF is lower than the BIF.

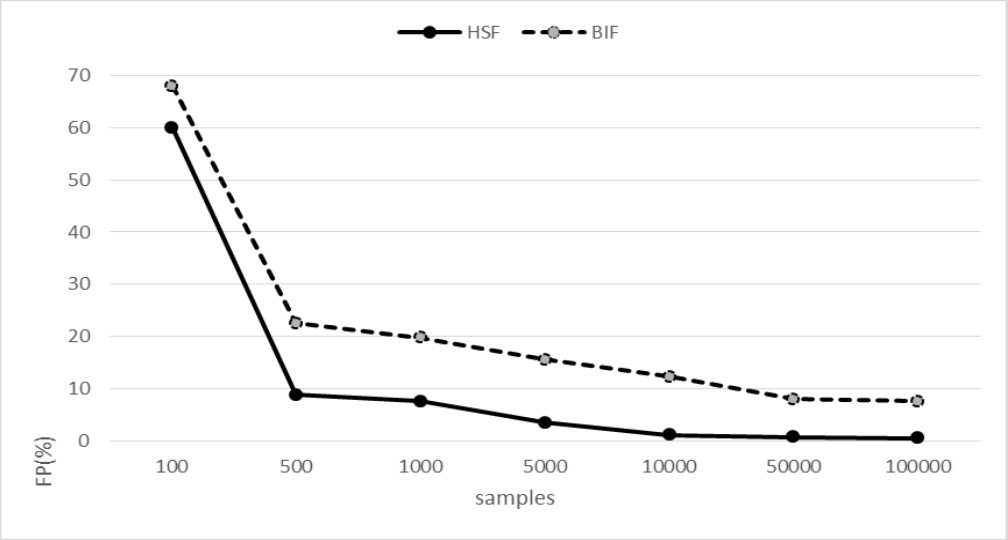


Figure 2. HSF and BIF Feature Selection Algorithm to Deal with the Same Amount of Data after the C4.5 Algorithm Classification FP Comparison

Figure 3 shows that with the increase of sample size, the memory of HSF and BIF algorithm in the feature selection process tends to rise. The memory usage of HSF algorithm is always higher than BIF, and the difference between them is obvious, mainly because during the initialization of HSF algorithm, the different characteristics of all permutations are stored into memory to be selected, but BIF only store each feature as feature selection into the memory, so HSF memory is always larger than the BIF. Observing From an upward trend, the memory occupancy of HSF rises slightly slower, a slight increase in the BIF rise in the latter part. This situation may because algorithm's processing speed is lower than the data read speed during the process and results in that with the increase amount of data, the data in the memory becomes much more. But both HSF and BIF algorithms are incremental feature selection algorithm, which is independent of the number of memory and data sample in theory.

 Copyright ⓒ 2015 SERSC

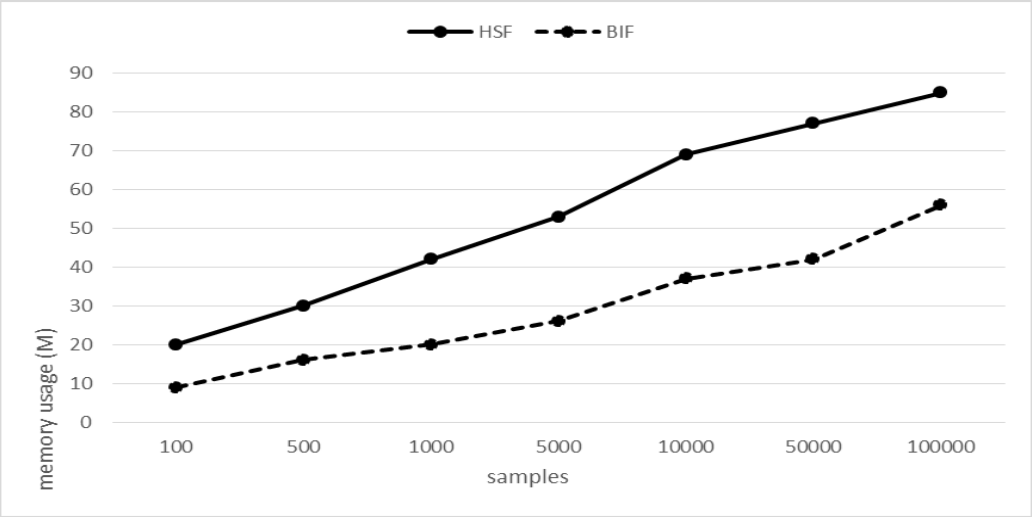


Figure 3. HSF and BIF Memory Usage Comparison with Different Samples

Figure 4 demonstrates that the processing time of HSF and BIF algorithm tend to rise with the increasing amount of data, because the processing rate is lower than the read rate while the sample size is always increasing, so the algorithm in the process of processing time is too long. The processing time of HSF is longer than that of BIF, and there is a small increase with the increase of sample size, but the gap between the two is almost constant, because the initial HSF calculations includes all permutations, but the BIF calculation only includes all the features so the amount of data processed by the HSF is greater than BIF. However, when the processing rate and the read rate reaches a certain point, the difference between the HSF and the BIF processing time also reaches a stable point.

Experiment-1 Results Conclusion: According to the results above, the performance of HSF is better than BIF in both TP and FP rate comparison. However because of the different constructing way of algorithm, the memory usage of HSF is much more than the memory usage of BIF. These mean that the accuracy of describing dataset obtained from HSF is better than BIF. HSF performs better than BIF excluding memory limit.

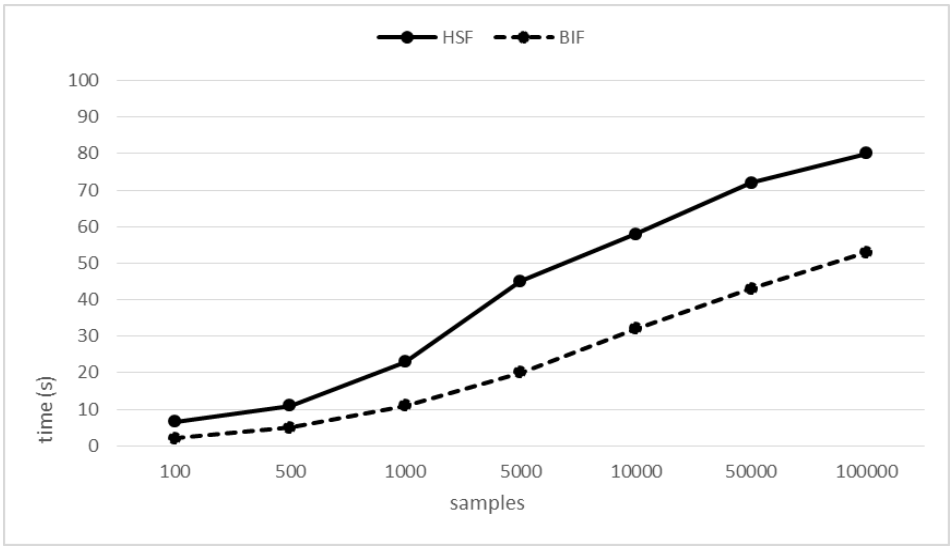


Figure 4. HSF and BIF Processing Time Comparison with Different Samples

Copyright ⓒ 2015 SERSC

Experiment-2: Figure 5 show that the feature subset selected by HSF algorithm has an obvious effect on improving the accuracy of classification. The feature subset of HSF selection can improve the classification accuracy, and the accuracy of the classification is improved with the increase of sample size, however, the feature subset of HSF selection is still outstanding in the accuracy rate.

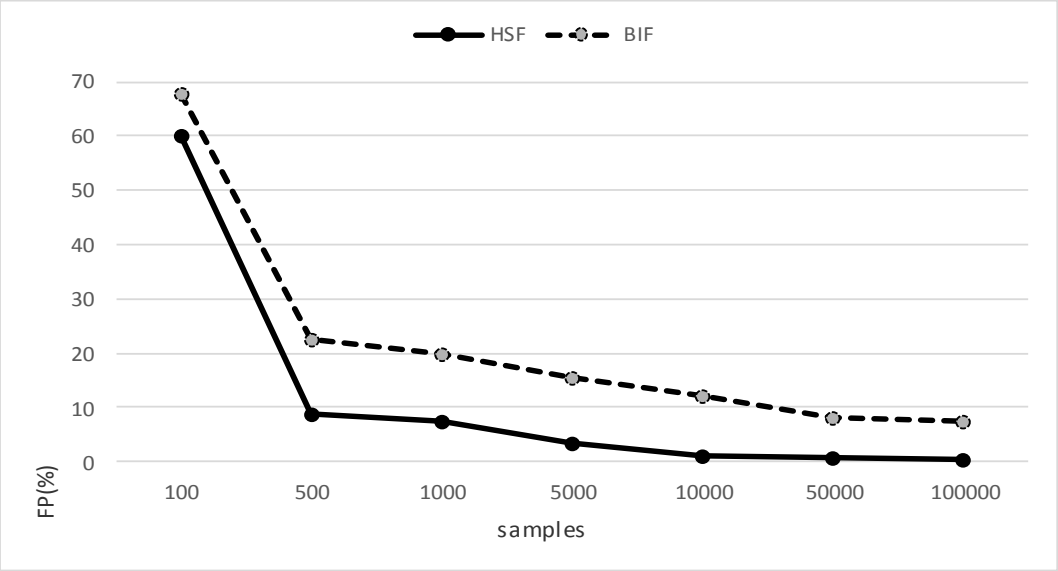


Figure 5. C4.5 Algorithm TP Comparison between Original Dataset and Subset Processed by HSF

Figure 6 shows that the feature subset selected by HSF algorithm has a large effect on reducing the classification error rate. The feature subset of HSF selection can effectively reduce the classification error rate. With the increase of sample size, the FP rate of the feature subset obtained by HSF algorithm is gradually reduced, and the difference between the two is very small.

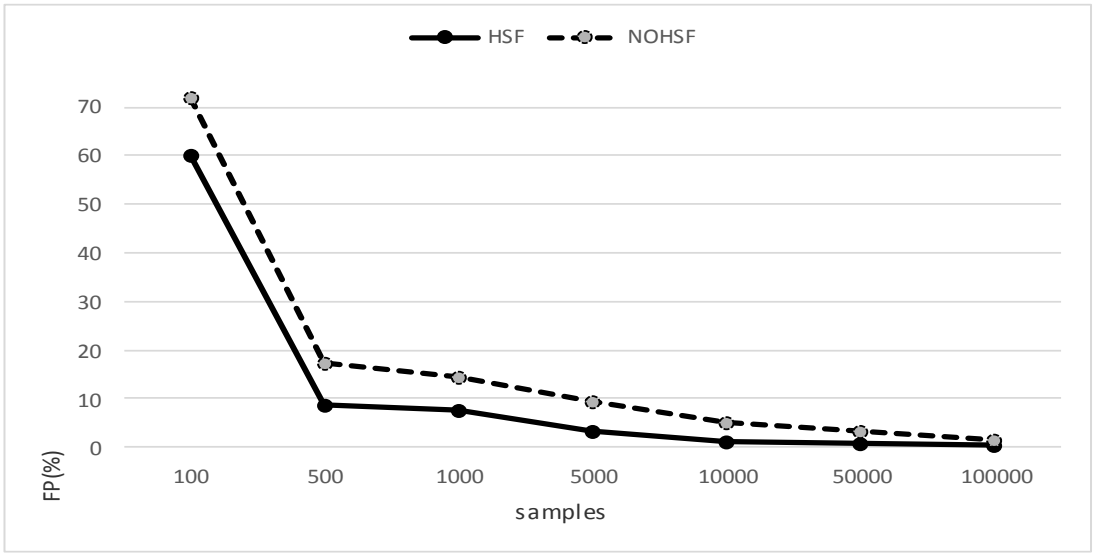


Figure 6. C4.5 Algorithm FP Comparison between Original Dataset and Subset Processed by HSF

Copyright ⓒ 2015 SERSC

Figure 7 show that the feature subset selected by HSF algorithm has a large effect on reducing the memory usage. The feature subset of HSF selection can effectively reduce the classification memory usage. With the increase of sample size, the difference of the feature subset obtained by HSF algorithm gradually tends to rise.

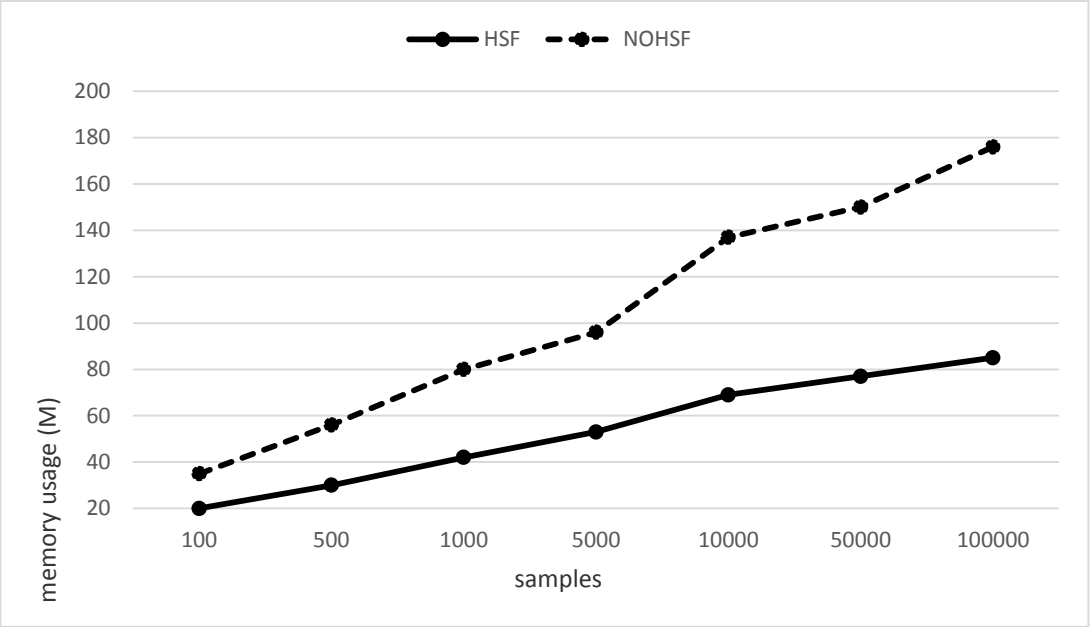


Figure 7. C4.5 Algorithm Memory Usage Comparison between Original Dataset and Subset Processed by HSF

Experiment-2 Results Conclusion: We can see the improvement of HSF applied to the classification algorithm. The subset obtained from HSF describes the dataset precisely, besides that HSF also has an outstanding performance on TP, FP and memory usage. These demonstrate that HSF algorithm can improve the efficiency of classification process.

5. Conclusions

In this paper, the basic concept of HSF algorithm based on mutual information is proposed, which is based on the set unit mutual information as a measure of the degree of association between different sets of different features. We take set unit mutual information as a measure of the association degree of the set of features and other collections. During the data process, Hoeffding inequality are introduced as compared to the termination condition to get rid of the previous single magnitude compared and avoid the data noise or bad data distribution influence the experimental results. Experimental results show that the classification accuracy and error rate of C4.5 classification algorithm are superior to those of the original feature set. In addition, with the comparing between the HSF and BIF based on mutual information, HSF have large advantages over BIF in the detection accuracy and detection error rate. But during the study, we find that all of the possible permutations and combinations are considered when the HSF algorithm is initialized, and it is placed in memory which results in memory occupancy and high cost of processing data. These problems above are also the author in the next step of the research need to improve the direction.

Copyright ⓒ 2015 SERSC

Acknowledgments

This paper is a revised and expanded version of a paper entitled “A feature selection algorithm based on set unit mutual information” presented at AITS 2015, Harbin, China, August 21-23, 2015. This work was funded by the National Natural Science Foundation of China (61373134, 61402234), and by the Industrial Strategic Technology Development Program (10041740) funded by the Ministry of Trade, Industry and Energy (MOTIE) Korea. It was also supported by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD), Jiangsu Key Laboratory of Meteorological Observation and Information Processing (No.KDXS1105) and Jiangsu Collaborative Innovation Center on Atmospheric Environment and Equipment Technology (CICAEET). Prof. Jin Wang is the corresponding author.

References

[1] J. Liu and G. Wang, "A hybrid feature selection method for data sets of thousands of variables", Advanced Computer Control (ICACC), 2010 2nd International Conference, **(2010)**.
[2] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering", Knowledge and Data Engineering, IEEE Transactions, vol. 17, no. 4, **(2005)**.
[3] K. Kira and L. A. Rendell, "A practical approach to feature selection", Proceedings of the ninth international workshop on Machine learning, **(1992)**.
[4] I. Kononenko, "Estimating attributes: analysis and extensions of RELIEF", Machine Learning: ECML-94, **(1994)**.
[5] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF", Machine learning, vol. 53, **(2003)**, pp. 1-2.
[6] Y. Sun and J. Li, "Iterative RELIEF for feature weighting", Proceedings of the 23rd international conference on Machine learning, **(2006)**.
[7] Y. Sun, S. Todorovic and S. Goodison, "Local-learning-based feature selection for high-dimensional data analysis", Pattern Analysis and Machine Intelligence, IEEE Transactions, vol. 32, no. 9, **(2010)**.
[8] C. Yin, "Towards Accurate Node-Based Detection of P2P Botnets", The Scientific World Journal, **(2014)**.
[9] C. Yin, M. Zou, D. Iko and J. Wang, "Botnet Detection Based on Correlation of Malicious Behaviors", International Journal of Hybrid Information Technology, vol. 6, no. 6, **(2013)**.
[10] B. Gu, V. S. Sheng, K. Y. Tay, W. Romano and S. Li, "Incremental Support Vector Learning for Ordinal Regression", **(2014)**.
[11] M. Bazarganigilani, "Web Service Selection Using Quality Criteria and Trust Based Routing Protocol", vol. 6, **(2012)**.
[12] B. Gu, V. S. Sheng, Z. Wang, D. Ho, S. Osman and S. Li, "Incremental learning for ν-Support Vector Regression", Neural Networks, vol. 67, **(2015)**.
[13] S. Park, D. Kyu Kim and B. Rae Cha, "Text Clustering using Semantic Features for Utilizing NFC Access Information", vol. 7, no. 3, **(2013)**.
[14] M. Hall, "Correlation Based Feature Selection for Discrete and Numeric Class Machine Learning", Proc. 17th International Conference of Machine Learning, **(2000)**.
[15] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy", The Journal of Machine Learning Research, vol. 5, **(2004)**.
[16] B. Guo, R. I. Damper, S. R. Gunn and J. D. Nelson, "A fast separability-based feature-selection method for high-dimensional remotely sensed image classification", Pattern Recognition, vol. 41, no. 5, **(2008)**.
[17] S. R. Kumari and P. Krishna Kumari, Adaptive Anomaly Intrusion Detection System Using Optimized Hoeffding Tree, **(2006)**.
[18] P. Domingos and G. Hulten, "Mining high-speed data streams", Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, **(2000)**.
[19] J. Wang, J.-U. Kim, L. Shu, Y. Niu and S. Lee, "A distance-based energy aware routing algorithm for wireless sensor networks", Sensors, vol. 10, no. 10, **(2000)**.

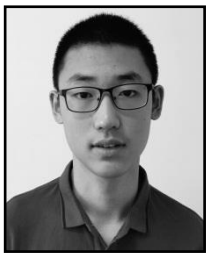
 Copyright ⓒ 2015 SERSC

Authors

Chunyong Yin, is currently an associate Professor and Dean with the Nanjing University of Information Science & Technology, China. He received his Bachelor (SDUT, China, 1998), Master (GZU, China, 2005), PhD (GZU, 2008) and was Post-doctoral associate (University of New Brunswick, 2010).He has authored or coauthored more than twenty journal and conference papers. His current research interests include privacy preserving and network security.

Lu Feng, received his bachelor degree in 2013 from Nanjing University of Information Science & Technology. His research interests are data-stream classification and feature extraction algorithm.

Luyu Ma, she received her BE degree in network engineering from Nanjing University of Information Science & Technology, China, in 2013. Currently she is a graduate student at the School of Computer and Software of Nanjing University of Information Science & Technology. Her research interests are in network security and intrusion detection.



Zhichao Yin, is studying in Nanjing No.1 Middle School. His current research interests include network security and mathematical modeling.

Jin Wang, received the B.S. and M.S. degree from Nanjing University of Posts and Telecommunications, China in 2002 and 2005, respectively. He received Ph.D. degree from Kyung Hee University Korea in 2010. Now, he is a professor in the Computer and Software Institute, Nanjing University of Information Science and technology. His research interests mainly include routing method and algorithm design, performance evaluation and optimization for wireless ad hoc and sensor networks. He is a member of the IEEE and ACM.

Copyright ⓒ 2015 SERSC

Copyright ⓒ 2015 SERSC