# Selection of online Features and its Application

Adhav Pradip N. 1
Computer Department,
SND college of Engineering,
Yeola,Pune University.
*pradipadhav33@gmail.com*

Bhalerao Amol R. 2
Computer Department,
SND college of Engineering,
Yeola,Pune University.
*amolrbhalerao96@gmail.com*

Deore Tushar J. 3
Computer Department,
SND college of Engineering,
Yeola,Pune University.
*tushardeore1993@gmail.com*

Shinde Nitin M.4
Computer Department,
SND college of Engineering,
Yeola,Pune University.
*ni33shinde29@gmail.com*

Guided By: Prof. Shaikh I.R., Professor of Computer Engineering, SND college of Engineering Yeola, SPPU, Maharashtra, India

***Abstract*—** Selection of Online Feature is significant important concept in data mining. Batch learning is the mostly used learning algorithm in feature selection. Instead of Batch learning, online learning is most efficient and scalable machine learning method. Most existing system studies of online learning should access the data related to features. But accessing all data becomes a problem when we deal with high dimensional data. To avoid this limitation we proposed system in this online learner allowed to operate a classifier having fixed and small number of features related data. But the significant challenge Selection of online features (SOF) is how to construct accurate prediction for a data using a small number of operative features. To develop novel Selection of Online Feature algorithms to perform a various tasks of Selection of Online Feature by using semi supervised and supervised with unlabeled and label data for full input and partial input. Hence it provides integrity and scalability to the data storage system efficiently and users will be accessing the data through online.

***Keywords****- Feature selection, online learning, large-scale data mining, classification, big data analytics.*

*****

## I. INTRODUCTION

Selection of online feature (SOF) is a significant important topic in data mining and machine learning and has been extensively studied for many years in literature [5]. Selection of online features (SOF) is an imperative step in successful data mining applications, which can effectively curtail data dimensionality by removing the irrelevant features. The intent of Selection of online features (SOF) is to select a subset of related features for building effective prediction models. By removing irrelevant, or noisy data, and repeated features, SOF algorithms can improve the performance of prediction models by subsidence the effect of the dimensionality, enhancing the generalizations performance, speeding up the learning process, and improving the model interpret ability [5]. In the past some decades, researchers have developing a huge number of SOF algorithms. These methods are designed to serve different purposes, of different module or methods, and all have their own merit and demerit.

Feature selection, a process of selecting a subset of original features and application according to specific criteria is an important and frequent used dimensionality reduces technique for data mining. SOF method to select a fixed and small number of features classification in an on-line learning fashion.

Most existing studies of Selection of online features (SOF) are limited to batch learning, which assumes that the Selection of online features (SOF) task is conducted in batch learning fashion and all the features of training instances are given advance. That assumption may not be always hold for real-world applications in which training instances arrived sequential manner or it is difficult to collect the expensive full information of training data [5]. Example is Selection of online features (SOF) is e-commercial website, where acquiring the entire set of features for every features is expensive due to the high cost.

In this paper, we address two different types of Selection of Online Feature system tasks: 1) Selection of Online Feature by learning with full inputs 2) Selection of Online Feature by learning with partial inputs. For the first task, we assume that the learner can access all the features of training data, and our goal is to efficiently identify a fixed number of most relevant features for accurate prediction. In the second task, the learner is allowed to access a fixe and small features for each training data to identify the subset of relevant features. To make this problem attractable, we allow the learner to decide which subset of features to acquire for each training data access through online.

## II. LITERATURE SURVEY

Our work is closely related to the studies of online learning and features selection in literature [5].

Prof. S.S Sane [2] has proposed dimensionality reduction techniques. In this system components work in two phases. In phase 1 the process of subset generation and evaluation is repeated up to the given criteria is satisfied. Subset generation element will produces the candidate features on the basis of search strategy. In phase 2 selected subset results is validate using two learning algorithms: 1) Online learning, 2) Online learning methods.

Steven C.H Hoi [3] has proposed the technique for large-scale application. In this system represent the effective algorithms to solve the empirical performance for OFS on several public datasets. In that system he demo the application of OFS technique for outfit real-world problem also evaluates the efficacy and efficiency of proposed system. In this also evaluate the online learning performance and online v/s batch learning comparison for real-world application based on this evaluation to checks the SOF results.

Divya K, Rama.B [1] has proposed OFS for gene selection. In this main aim is to reduce the dimensionality of of microarray dataset. In this gene finding perform the class comparison to determine the quantities measurements and survival time. it also represents the full and partial input features selection methods to reduce the search dimension of data.

N. Cesa-Bianchi [6] has proposed the integrity and scalability to data storage system .In this system only partial input is provided to accessing the data through online. The partial input is given to the features selection process system perform the features selection in three diffract process supervised, un-supervised and semi-supervised in supervised process it uses the three method wrapper, filter and embedded.

## III. Proposed system

Selection of online features which aims to select a small and fixed number of features for classification in an online learning fashion [5]. To develop novel Selection of Online Feature approaches which are compared with existing classification algorithms and to measure its performance for real-world datasets with full and partial inputs? We presented a family of novel Selection of Online Feature algorithm to solve each & every of the Selection of online features (SOF) tasks, and offered theoretical analysis on the mistake fixed in the proposed SOF algorithms.

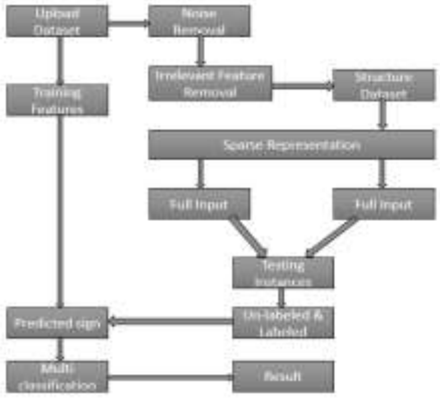


**Fig -1**: System architecture.

In this process first upload datasets & then Perform pre-processing steps to remove irrelevant features from uploaded datasets. Data pre-processing includes data cleaning, data integration, data transformation and data reduction [7].Then the sparse representation for both full and partial input. Implement sparsely property to overcome misclassified features. Construct training instances to get full knowledge features. And analyze both labeled and unlabeled instances from users. The features are already trained in the data sets. Predicted Sign is used to match classifier with testing instances each testing instance has corresponding training instances. Multiclass classification algorithm to analyze sub features. Then finally find the prescribed results.

.

### A. Problem Statement

In this section, we present a novel Selection of Online Feature method. We first describe the problem setting and then introduce a SOF algorithm, followed by representing the proposed method in detail.

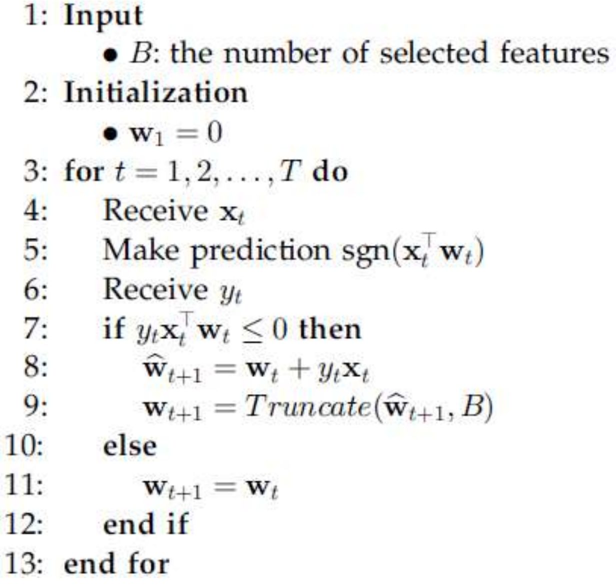
### B. Learning with Full Input

To motivate our algorithm, we first present a simple but no effective algorithm that simply truncates the features with small weights. The failure of this simple method motives we are develop an effective algorithms for SOF.

### C. A Truncation Approach

In truncation approach to Selection of Online Feature is to modify the perception algorithm by applying truncation. Specifically we will truncate the classifier wt by setting everything but the B largest elements in wt. to be zero. This truncated classifier, denoted by wB, Then used to classify the receive example xt Similar to the Perception algorithm, when the example is misclassified, we will update the classifier by adding the vector yt xt where (xt, yt ) is the misclassified training example. This simple approach does not work it cannot guarantee a small number of mistakes [5].

*Algorithm 1: Modified Perceptron by Truncation for SOF*



```
Input
    • $B$: the number of selected features
Initialization
    • $\mathbf{w}_1 = 0$
for $t = 1, 2, \ldots, T$ do
    Receive $\mathbf{x}_t$
    Make prediction $\text{sgn}(\mathbf{x}_t^\top \mathbf{w}_t)$
    Receive $y_t$
    if $y_t \mathbf{x}_t^\top \mathbf{w}_t \leq 0$ then
        $\widehat{\mathbf{w}}_{t+1} = \mathbf{w}_t + y_t \mathbf{x}_t$
        $\mathbf{w}_{t+1} = Truncate(\widehat{\mathbf{w}}_{t+1}, B)$
    else
        $\mathbf{w}_{t+1} = \mathbf{w}_t$
    end if
end for
```

### D. A Sparse Projection Approach

Truncation approach does not work well. It cannot guarantee little mistakes as it doesn't assurer the numerical values of truncated elements are leading to a loss of accuracy. This, they propose a first-order Selection of Online Feature scheme by exploring online gradient descent with a sparse projection scheme which guarantees the resulting classifier wt. to be restricted into a 1 -ball at each step.

Selection of online features (SOF) algorithm via the sparse projection. They also give a theoretical analyze and shows the bounded weight vector will lead to the bounded number of mistakes. Howsoever, this algorithm in general has a linear time complexity with respect to the feature dimensionality, which could be slow for high dimensional data, and its empirical Performance might not be always satisfying when handling difficult Selection of online features (SOF) tasks.

Most of its numerical values are concentrated in its largest elements, and therefore removing the smallest elements will result in a very small change to the original vector measured [5]. The theorem gives the mistake bound of Algorithm Based on this idea; we present a new approach in Algorithm for Selection of Online Feature. The online learner maintains a linear classifier wt that has at most B non-zero elements. When a training instance (xt, yt) is misclassified, the classifier is updated by online gradient descent and then projected; since after projection, the norm of the classifier could be bound, which is inspired by the Pegasus algorithm [5].

### E. Learning with Partial Inputs

It provides tradeoff between exploration and exploitation. In this approach, to spend ε of trials for exploration by random choose B attributes from all d attributes, and the remaining 1−ε trials on exploitation by choosing the B attributes for which classifier withes non-zero values. Partial input contains both labeled and unlabeled data. Labeled data means class label is defined and unlabeled data means the class label is not defined but it is based on similarity between data. And implement predicted sign to create new data elements based on labeled or unlabeled testing instances. Predicted sign is used to predict the patterns corresponding to the given input. Labeled data works in supervised manner, which define the class label in the group of patterns.

It is easily identifying the particular pattern in the group because the class label is defined. So implement the predicted sign to predict the particular sign in group of gene patterns corresponding to the given input. Unlabeled data works in unsupervised manner, which not define the class label. This approach works with similarity between data in pattern. The input data is taken and it is matched for all other data in pattern group. If matched pattern is added to the new data element and unmatched elements are removed.

## CONCLUSION

We study different approaches for Selection of online features (SOF) and efficiency of SOF algorithm against these approaches. Selection of online features which aims to select a fixed and small number of features for classification. In SOF we discuss two feature selection tasks in first learning with full inputs and second task learning with partial input.

## REFERENCES

[1]. Divya.K, Rama,"Interactive Supervised Feature Selection Framework In Microarray Datasets", ICETSH-2015.

[2]. Prof. Dr. S. S. Sane**,** M. Tech (CSE, IITB), Ph D (COEP, Univ. of Pune), Vice Principal, Professor & Head of Dept. of Computer Engineering, Prof. In-charge Central Library. K. K. Wagh. Volume 3, Issue 10, October 2014.

[3]. Jailed Wang, Peilin Zhao, Steven C.H. Hoi, Member, IEEE, and Rong Jin, "Online Feature Selection and its Applications," IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 3, March 2014.

[4]. Hala M. Alshalan, Ghada H, Badt, and Yousefalohali a study of cancer microarray gene expression profile: objectives and approaches proceedings of the world congress on engineering 2013 Vol II, WCE 2013, July 3-5, 2013, London, UK.

[5]. Hoi, Steven C. H., Jialei Wang, Peilin Zhao, and Rong Jin. "Online Feature Selection for mining big data", Proceedings of the 1st International Workshop on Big Data Streams and Heterogeneous Source Mining Algorithms Systems Programming Models and Applications - Big Mine 12 Big Mine 12, 2012

[6]. X. Wu, K. Yu, H. Wang, and W. Ding, "Online Streaming Feature Selection," Proc. Int'l Conf. Machine Learning (ICML '10), pp. 1159- 1166, 2010.

[7]. Zhang, Tiefeng, Jie Lu, Guangquan Zhang, and Qian Ding. "Fault diagnosis of transformer using association rule mining and knowledge base", 2010 10th International Conference on Intelligent Systems Design and Applications, 2010.

[8]. J. Ren, Z. Qiu, W. Fan, H. Cheng, and P. S. Yu.Forward semi-supervised feature selection. In PAKDD, pages 970–976, 2008.

[9]. Y. Saeys, I. Inza, and P. Larranaga, "A Review of Feature Selection Techniques in Bioinformatics," Bioinformatics, vol. 23, no. 19,pp. 2507-2517, 2007.

[10].Z. Zhao and H. Liu, "Spectral Feature Selection for Supervisedand Unsupervised Learning," Proc. Int'l Conf. Machine Learning (ICML '07), pp. 1151-1157, 2007.

## BIOGRAPHIES

Adhav Pradip N perceiving the B.E. degree in Computer Engg. From S.N.D COE & RC, Yeola in 2015.

Bhalerao Amol R perceiving the B.E. degree in Computer Engg. From S.N.D COE & RC, Yeola in 2015.

Deore Tushar J perceiving the B.E. degree in Computer Engg. From S.N.D COE & RC, Yeola in 2015.

Shinde Nitin M perceiving the B.E. degree in Computer Engg. From S.N.D COE & RC, Yeola in 2015.