Classification using Online Feature Selection with Partial Inputs

Deshmukh Ankita Chandrabhan **Prof. Dr. S. S. Sane**

Abstract — **Feature Selection has been an active field of research topic that plays a significant role for decades in machine learning, and data. For quite some time, feature selection has remained as a focus of interest and even much work has been reported in this area. Due to new rising problems in the case of huge databases new approaches to feature selection are in demand. Mostly, feature selection research focuses on batch learning. For large scale applications, online learning is also promising as well as adequate and scalable machine learning algorithms. Most of the studies of online learning requires to access all the features of training instances. Such a classical setup is not always applicable for real-world applications of high dimensional data instances or it is expensive to achieve the full set of features. To overcome this limitation, researchers have investigated the online feature selection (OFS) problem in which an online learner is allowed to maintain a classifier along with only small and fixed number of features. Making accurate prediction for an instance by using a small number of active features is the key challenge of online feature selection. This is distinct from classical setup of online learning wherein all the features can be used to predict.**

This research aims at implementation of an algorithm for online feature selection and to analyze its performance for real-world datasets with full and partial inputs. The work also aims at investigating the performance of the algorithm with multi class datasets.

Keywords - **Classification, feature selection, high dimensional data mining, online learning**

I. INTRODUCTION

Computer field is growing very rapidly from 21st century. Many advance techniques and technologies are developing which has lead to a great progress and increased opportunities for peoples to increase research, communication, production and services. Due to this growth quantity of data being used is also increasing day by day. Thus, working on such a high dimensional data and maintaining the quality is a big challenge which is tried to be achieved by latest data mining methods. Selection of effective feature i.e. removing irrelevant and duplicate features is a necessary task in successful data mining applications. In last few decades researchers have developed large amount of feature selection algorithms. These algorithms have different purpose and also have their own advantages and disadvantages. Feature selection means selection of subset of features from the set of original ones by following some criteria. It is a dominant and commonly used technique for dimensionality reduction in data mining. Thus it quickens the mining algorithm, and also improves its efficiency. For classification, the goal of feature selection is to select a subset of relevant features to build adequate prediction models. If good technique is used for selecting features than it speeds up the performance of prediction models by reducing the effect of the curse of dimensionality, increasing the speed of the learning process, improving the generalization performance and also the model interpretability. The feature selection technique is used in many applications viz. for reducing the problems related to big data and for dimensionality reduction technique.

Though feature selection is used as well as studied widely in many domains, the methods or techniques used are mostly associated with batch learning methods. Offline learning is other name given for batch learning. In batch fashion, at the time of training phase all the features of training instances are taken into consideration. But in actuality for high dimensional data or when instances come sequentially, considering all features of training instances may not be applicable. It may be expensive too. For example, if we consider online spam detection system than training data usually arrive sequentially due to which using batch selection in such cases in efficient, scalable, and timely way is difficult. The aim of proposed system is to develop OFS algorithm for training the classifier as discussed in [1]. OFS depends on several factors, such as, feature extraction method, suitable features to be used, calculation of mistake rates etc. All these factors are important in OFS, since an improvement to any of these influencing factors can result in a more effective classifier training mechanism. With this view in mind, the aim is to improve the performance of existing learning algorithms. OFS algorithm is the most effective algorithm used for binary classification. But for multiclass classification OFS method is not yet implemented. Thus, the main focus is on the implementation of OFS for multiclass classification using partial inputs. Also, performance evaluation is to be carried out with multiclass datasets.

II. LITERATURE SURVEY

In this section, some of the existing feature selection techniques are overviewed. The objective of this survey is to clearly understand the limitations of existing schemes.

One of the most popular feature selection algorithm is the Perceptron algorithm [2]. Recently, several online learning algorithms have been introduced in which most of them work on the basis of classification margin range principle method. For instance, in the Passive-Aggressive (PA) algorithm [6] if the incoming training data is either not classified properly or it falls into the classification margin range then the classifier needs to be updated. The PA algorithm follows an aggressive update method i.e. 1st the weight vector is modified until a specified constraint level is met. But in real life it may lead to certain problems. For example, if something is mislabeled than it may cause the weight vector to change drastically in other direction or it may lead to prediction mistakes. To resolve such type of problems, the PA uses moderate update strategies. In spite of the comprehensive research, most studies of online learning require the access to each and every feature of training data. On the contrary, we consider an online learning problem. In this problem the learner is only allowed to access a small and fixed number of features. This is more challenging problem as compared to the traditional system of online learning.

Feature Selection has been widely studied in the literatures of data mining and machine learning [4], [5]. The FS algorithms can be generally grouped into 3 categories: unsupervised FS, supervised feature selection, and semi supervised FS as shown in fig. 1. Supervised FS selects features according to labeled training instances. Supervised selection is further classified into 3 models viz. embedded model, filter model, and wrapper model. In filter method, features or groups of features are scored by some measure of correlation with the labels. Fisher scores, mutual information are well known examples of scoring functions. Wrapper model uses the classifier to guide the process of selection. Embedded model selection method uses properties of the classifier. If the labeling is not available, unsupervised feature selection is used. Unsupervised FS selects the important features which preserve the original data similarity. Feature selection is widely used in the fields of [5], text analysis, bioinformatics, and image annotation [7]. Finally, in recent years some semi-supervised feature selection methods are also introduced that uses unlabeled as well as labeled data information [8], [9], [11]. The Online Feature Selection technique comes under the supervised feature selection category.

Dimensionality reduction through sparse vector machines [3] is closely related to sparse online learning. Sparse online learning method from a sequence of high-dimensional training examples aims to learn a sparse linear classifier. The proposed work however differs from these studies as it explicitly addresses the feature selection issue and thus enforces a hard constraint on the number of nonzero elements in classifier. Most of the previous studies of sparse online learning do not explicitly address feature selection, and usually impose only soft constraints on the sparsity of the classifier. In spite of the difference between OFS and sparse online learning, the proposed online feature selection algorithm performs better than the sparse online learning algorithms for online classification tasks when the same sparsity level is imposed for the two algorithms.

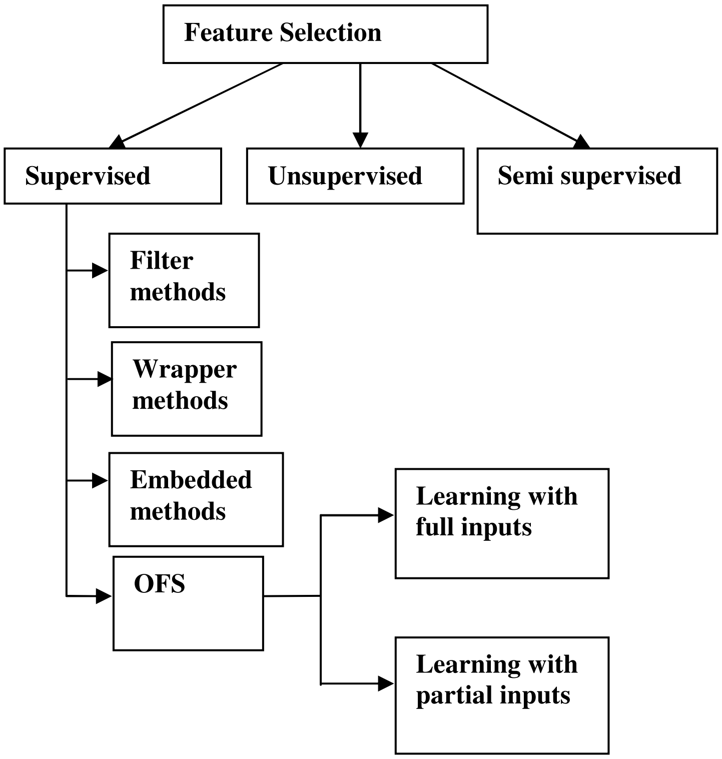

Fig.1. Classification of Feature Selection Techniques

In online streaming works [6], features are expected to arrive one at a time and also before the learning process starts, all the training instances are assumed to be available and aims to train an appropriate model by selecting a subset of features at each time step. Our online learning setting significantly differs from this where training instances arrive sequentially, a more natural scheme in real-world applications.

Online feature selection for mining big data [10] introduces the OFS algorithm for learning with only full inputs. Learning with full inputs becomes complex for high dimensional data and thus time required for processing is also large. In the proposed work we are using online feature selection with partial inputs which gives the maximum accuracy as compared to previous algorithms.

III. PROPOSED WORK

A. Proposed System:

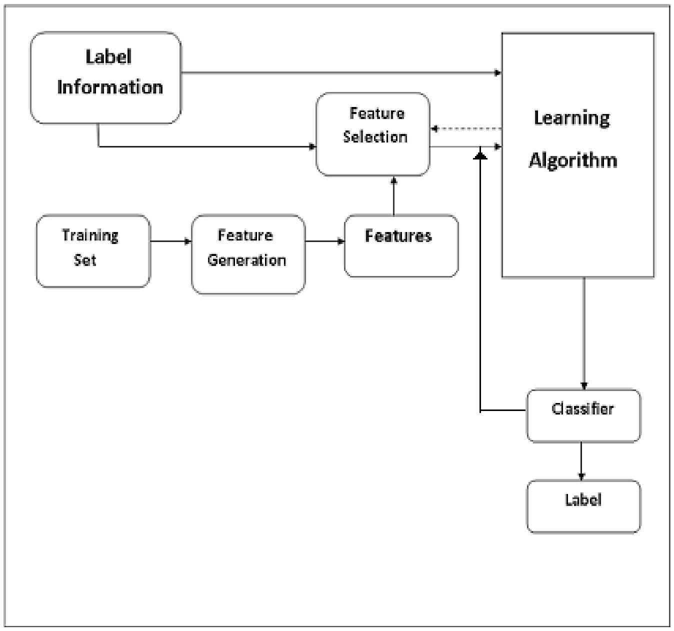

Fig. 2. Block diagram of proposed system

The above fig. 2 shows the block diagram of the proposed system. The training and prediction phase are merged in online learning process. Firstly the training set i.e. dataset is taken as an input.

From the dataset the first instance is extracted. From the extracted instance active features are stored in features database and other features are truncated. Active features are the positive or the nonzero values.

From the active features randomly some features are selected called as the partial features and with the label information it is given as an input to the learning algorithm which trains the classifier.

Than if ever classifier makes mistake the counter is incremented, weight vector is updated and is again given as input to the learning algorithm with the partial features of next instance. This process is repeated sequentially for all instances of the respective dataset for n number of iterations.

At last the average number of mistakes made by our system is calculated and on its basis the performance is measured.

IV. Experimental Setup

All Experimentation is performed using Pentium processor and 4 GB RAM. The operating system is windows 7(32 bit) with c#.net coding.

A. *Dataset:*

The data sets used are numeric dataset both for binary and multiclass classification. The experimentation shown in base paper is approximate to our base paper implementation. **Spambase dataset** is used for binary classification.

For our proposed system i.e. multiclass classification **KDDCUP99, IRIS** multivariate datasets are used.

B. *Performance Measure:*

Performance of the algorithm is calculated with the **average number of mistakes** made while training the classifier. The less the number of mistakes the more is accuracy of the system..

C. *Results:*

Table I. shows that the experimentation of the base paper results is approximate to our results of base paper.
Analysis is shown for Spambase dataset.

Algorithm	Results shown in base paper	Implementation of base paper
RAND	3278.6 +- 40.4	3248.3+- 67.5
PE_{trunc}	3275.4 +- 40.4	3251.4 +- 67.5
OFS	1954.2 +- 78.7	1958.9 +- 80.4

Table I. Evaluation of the Average Number of Mistakes by Three algorithms on Spambase Datset

From Table I. we can see that OFS is efficient and performs better than other two algorithms for binary classification. But its efficiency in case of multiclass classification was not yet measured.

Thus, proposed system proved that when implemented the OFS algorithm for multiclass classification it again gave better performance than RAND and PE $_{trunc}$.
Table II. shows the results for KDDCUP99 multiclass dataset.

Algorithm	KDDCUP99
PE	2845.5 +- 71.5
PE_{trunc}	2094.5 +- 79.1
RAND	2089.5 +- 79.1
OFS	1260 +- 87.4

Table II. Evaluation of the Average Number of Mistakes by Three algorithms on KDDCUP99 Dataset for proposed system

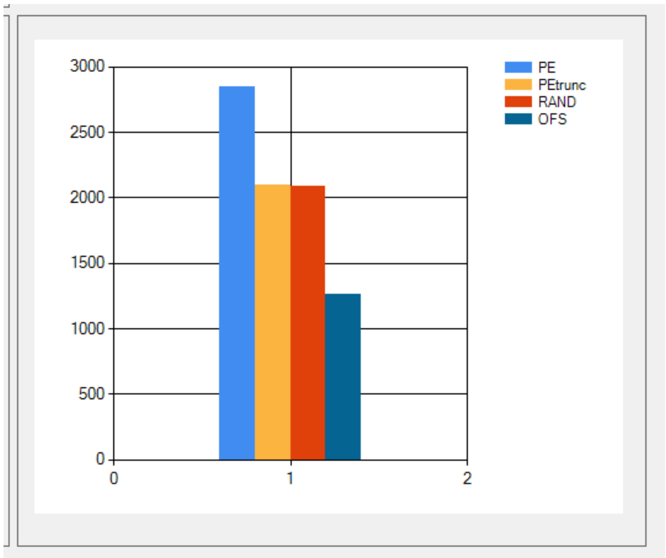

Fig. 2. Performance evaluation of online feature selection in the online learning process using multiclass classification.

Fig. 2 shows the details of online average mistake rates for the entire OFS process on the KDDCUP99 dataset. Similar to the previous observations, we can see that the proposed OFS algorithm outperformed the other algorithms.
This validates the effectiveness of the proposed technique.

CONCLUSION

Huge databases leads to many problems. Due to these new arising problems, new approaches to feature selection are introduced. One of the effective and scalable learning techniques is online learning. For large scale applications online learning is proven to be the best solution. Most of the studies of online learning accesses full set of features. But in case of high dimensional data accessing full set of features in real world application is too expensive. To overcome this, OFS is an alternative. OFS maintains a classifier with only small and fixed number of features.

OFS for binary classification has been implemented and is proven to be the better algorithm than classical binary classification algorithms. Thus, analyzing OFS for multiclass classification was a big challenge.

Thus, the proposed system aimed to analyze OFS for multiclass classification using partial input. When the OFS system for multiclass classification is experimented it outperforms same as the OFS for binary classification. This system also generates the mistake rate that helps us out for analyzing the performance of OFS multiclass classification with classical methods.

ACKNOWLEDGMENT

I would like to express my special thank to all those people who have helped me to complete this work. I am very grateful to my guide, Prof. Dr. S. S. Sane, Head of Department, Computer Engineering, KKWIEER, Nasik for his guidance, encouragement and the interest shown in this project. He has continuously helped and encouraged me in my work.

REFERENCES

[1] Jialei Wang, Peilin Zhao, and Steven C. H. Hoi, "Online feature selection and its Applications," IEEE transaction on knowledge and data engineering, vol. 26, no. 3, March 2014.

[2] F. Rosenblatt, "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain," Psychological Rev., vol. 65, pp. 386-407, 1958.

[3] J. Bi, K.P. Bennett, M.J. Embrechts, C.M. Breneman, and M. Song, "Dimensionality Reduction via Sparse Support Vector Machines," , j.Machine Learning Research, vol. 3, pp. 1229-1243, 2003.

[4] M. Dash and H. Liu, "Feature Selection for Classification,"Intelligent Data Analysis, vol. 1, nos. 1-4, pp. 131-156, 1997.

[5] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," J. Machine Learning Research, vol. 3, pp. 1157-1182, 2003.

[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] Z. Ma, Y. Yang, F. Nie, J.R.R. Uijlings, and N. Sebe, "Exploiting theEntire Feature Space with Sparsity for Automatic Image Annotation," Proc. 19th ACM Int'l Conf. Multimedia , pp. 283-292, 2011.

[8] Z. Zhao and H. Liu, "Semi-Supervised Feature Selection via Spectral Analysis," Proc. SIAM Int'l Conf. Data Mining (SDM '07),2007.

[9] J. Ren, Z. Qiu, W. Fan, H. Cheng, and P.S. Yu, "Forward Semi-Supervised Feature Selection," Proc. 12th Pacific-Asia Conf. Advances in Knowledge Discovery and Data Mining (PAKDD '08), pp. 970-976, 2008.

[10] S.C.H. Hoi, J. Wang, P. Zhao, and R. Jin, "Online Feature Selection for Mining Big Data," Proc. First Int'l Workshop Big Data, Streamsand Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications (BigMine '12), pp. 93-100, 2012.

[11] Z. Xu, I. King, M.R. Lyu, and R. Jin, "Discriminative Semi-Supervised Feature Selection via Manifold Regularization," IEEE Trans. Neural Networks, vol. 21, no. 7, pp. 1033-1047, July 2010.

AUTHOR'S PROFILE

Ankita Deshmukh, received the B.E. degrees in Information Technology from SVIT College of Engineering, Nashik, Savitribai Phule Pune University in 2013. Now pursuing M.E. in Computer Engineering from K. K. Wagh Institute of Engineering Education & Research, Nashik, India.

Prof. Dr.S.S.Sane received M. Tech (CSE) from IITB, Ph D from COEP, University of Pune. Currently working as Vice Principal, Professor & Head of Dept. of Computer Engineering, Prof. In-charge Central Library in K K Wagh Institute of Engineering Education & Research, Nashik, India**.**