International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 4 121 - 122

Review on OFS: Online Feature Selection based on Regression analysis and Clustering method along with its Application

Priyanka Vhansure[1]
[1]M.E.(CSE),
Department of Computer Science and Technology,
NBNSCOE,Solapur University,
Solapur, India
priyanka.vhansure@gmail.com

A. A. Phatak[2]
[2] Asst. Prof
Department of Computer Science and Technology,
NBNSCOE, Solapur University,
Solapur, India
amolphatak@rediffmail.com

Abstract-In Data mining the Feature selection is one of the main techniques. In this its result shows, almost all learning of feature selection is finite to batch learning. Not similar to existing batch learning methods, online learning can be chosen by an encouraging family of well-organized and scalable machine learning algorithms for large-scale approach. The large scale quantity of online learning needs to retrieve all the features/attributes of occurrence. The difficulty in Online Feature Selection in which the online learner is allowed to maintain a classifier that involved a small and fixed or exact number of features. This article demonstrates two different tasks of online feature selection. First one is learning with full input and second is learning with partial input. The sparsity regularization and truncation techniques are used for developing the algorithms. There is a challenge of online feature selection is how to make prediction accurately for an instance using a small number of active features in high dimensionality. The proposed system presents novel method such as Multiclass classification, Regression analysis and Clustering method to clear up each of the two problems and give their performance analysis.

Keywords:- *Feature Selection, Online learning, Classification, Regression, Clustering.*

*****

1. INTRODUCTION

Feature selection is an significance step in successful data mining applications. In Feature Selection process batch learning is incessantly used. It can be effectively reduce data dimensionality by discard the irrelevant and the redundant features. Feature selection, a process of selecting a subset of original features according to certain criteria, is an important and often used dimensionality reduction technique for data mining. It reduces the number of features, removes unrelated, redundant, or noisy data, and brings the immediate effects for applications: speeding up a data mining algorithm, and getting better mining performance such as predictive accuracy and result comprehensibility.

The goal of Feature Selection (FS) is to choose the most relevant features in the whole feature space to increase the prediction performance of the predictors. Feature selection is divided into 3 categories: filter, wrapper and embedded. The purpose of Online feature selection is to resolve the feature selection problem in an online fashion by effectively exploring online learning techniques.

Online learning needed all the attributes or features of training instance. The Online Feature Selection aims to select a less and fixed number of features for multiclass classification in an online learning fashion. OFS gives two different types of tasks in different settings:

First one task is OFS by learning with full inputs, in this task learner is allowed to access full features to decide the subset of active features, and second one task is OFS by learning with partial input, in this task learner is allowed to access only limited features for each instances.

Sparsity regularization and truncation techniques are used for implementing algorithms. These algorithm designed to use different purposes, and different model and also their own advantages and disadvantages. Feature selection has foundmany applications or uses inmany domains or fields, especially for the problems involved high dimensional data. Such assumptions may not always be applicable for real-world examples in which training examples arrive in order it is expensive to collect the all information of training data. A online spam email detection system is the example, in this training data usually arrive in order, making it requiring much effort to deploy a regular batch feature selection method in a timely, efficient, and scalable manner. Bioinformatics is another example of feature selection, where acquiring the entire set of features/attributes for each training instance is expensive due to the high cost for conducting experiments.

2. RELATED WORK

Some existing studies and approaches related for online feature selection to resolve the feature selection problem in an online fashion by effectively shows online learning techniques which are following.

2.1 ONLINE PASSIVE-AGGRESSIVE ALGORITHMS [1]

In this Passive-Aggressive algorithm uses to update a classifier when the incoming training example is either misclassified or drop into the range of classification margin. The PA algorithm is limited in that it only uses the first order information during the updating. For the second order information the confidence weighted online learning algorithm is used which is recently proposed and in this PA algorithms limitation has been addressed. Alternative modifications to the PA algorithm which enhance the algorithm's capability to cope with noise are proposed. A

unified analysis for the three variants is also proved implementing on this unified view, here is show how to generalize the binary scenario to various learning tasks, ranging to sequence prediction from regression.

2.2 DIMESIONALITY REDUCTION via SPARSE SUPPORT VECTOR MACHINES [2]

This work has presents its effectiveness on very high-dimensional problems with very small amount of data. On problems where linear models cannot sufficiently capture relationships, the method would fall. Open research areas include a theoretical foundation of the approach, characterization of the area on which it is effectual, and extension to nonlinear interactions.

2.3 ONLINE STREAMING FEATURE SELECTION [3]

In the online streaming feature selection, in this the size of the feature set is unfamiliar, and not all features are usable for learning while leaving the number of observation constant.

2.4 ONLINE FEATURE SELECTION FOR MINING BIG DATA [4]

In this presents the application of OFS method to deal with real-world problems of big data mining, which is way more scalable than some well-known batch feature selection algorithms.

2.5 LEARNING WITH MISSING FEATURE [4]

In this presents new online and batch algorithms that are strong data with missing features thiscondition arises in many practical software applications. In the online setup, the comparison hypothesis to update as a function of the subset of features that is observed on any given circular, extending the standard setting where the comparison hypothesis is fixed throughout.

2.6 ONLINE FEATURE SELECTION AND ITS APPLICATIONS [5]

This paper introduces online learning has two approacheswhich are implemented: 1) Learning with full inputs 2) Learning with partial inputs. For this Sparsity regularization and truncation these techniques are used for developing the algorithms.For the first approach, it is supposed that the learner can access all the features of training instances, and goal is to efficiently find exact number of relevant features for accurate and exact prediction. In the second task, a many challenging scenario is considered where the learner is allowed to access a fixed less number of features for each training instance to find the subset of significant features.

3.CONCLUSION

The Online Feature selection or OFS technique introduces the process of selecting a subset of relevant features which is used for building a software model. The goal of OFS is to select a small and fixed number of features by using classifier in an online learning fashion.OFS is mainly used for eliminating the same features in large-scale data the redundant and irrelevant data do not contribute to the accuracy of a predictive model or may in fact decrease the accuracy of the model. Small number of attributes are desirable because it reduces the complexity of the modeland a simple model is comfortablyunderstand and explain.

The OFS approach is mostly suitable for high dimensional data contains multiclass classification to tackle so many real-world problems. OFS based on regression analysis and clustering method it becomes fast learning process for selecting the features. It can improve the performance and complexity of the prediction model and also increases the accuracy of the model.

4. REFERENCES

[1] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer,“Online Passive Aggressive Algorithms,” J. Machine Learning Research, vol. 7, pp. 551-585, 2006.

[2] J. Bi, K. P. Bennett,M. J. Embrechts, C. M. Breneman, and M. Song.Dimensionality reduction via sparse support vector machines.Journal of Machine Learning Research, 3:1229–1243, 2003.

[3] X. Wu, K. Yu, H. Wang, and W. Ding,“Online Streaming Feature Selection,” Proc. Int’ Conf. Machine Learning (ICML ’10), pp. 1159 1166, 2010.

[4] 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, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications (BigMine ’12), pp. 93-100, 2012.

[5] “Online Feature Selection and Its Applications” , Jialei Wang, Peilin Zhao, Steven C.H. Hoi, Member, IEEE, and Rong Jin, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 3, MARCH 2014.

AUTHOR

[1] Priyanka R. Vhansurereceived the B.E. degree and pursuing M.E in Computer Science & Engineering from N. B. Navale Sinhgad College of Engineering Solapur, India.

[2] Prof. A A. Phatak working as Assistant Professor, Department of Computer Science and Engineering, N. B. N. Sinhgad College of Engineering, Solapur, India.