An Optimal Machine learning Approach for Large-Scale Data

P.Sirisha[1], M.V. Jagannatha Reddy [2]

[1]PG Student, Department of Computer Science and Engineering, Madanapalle Institute of Technology & Science, India.

[2]Assistant Professor, Department of Computer Science and Engineering, Madanapalle Institute of Technology & Science, India.

Abstract: Online learning is an important module of optimal machine learning techniques used in large-scale application developments. Feature selection is found in many applications to solve the problems raised by high dimensional data. Our goal is to develop an efficient prediction model through pruning the dataset by filtering the irrelevant and duplicates. It ensures the best performance in learning, interpretation, dimensioning .We propose online feature selection on the medical domain for predicting the treatment and side-effects on the diseases. We collect the dataset from a medical blog (midline).which is traversed between unsupervised, semi-supervised & the supervised .We introduce 2C & 3C classifications for feature selection, we train the system with the selected features and the result shows the treatments recommended by the medical experts in the online datasets.

Keywords*:* Feature Selection; Online Learning; SVM.

1. INTRODUCTION

The Feature Selection is an important area in data mining and machine learning, and has been extensively considered for lots of years. The purpose of attribute selection is to choose a subset of related features for structure efficient calculation models. By way of remove unrelated and unneeded features, FS can recover the show of calculation model by alleviate the outcome of the curse of dimensionality, striking the simplification presentation, speeding up the knowledge process, along with improving the model interpretability. Feature selection has set up applications in several domains, in particular for the problems concerned high dimensional data. Despite being studied widely, most presented studies of feature selection are controlled to batch learning, which assumes that the feature selection assignment is conducted in an offline/batch learning fashion and all the features of training instances are given a priori. Such statement cannot always hold for real world applications in which training examples appear in a sequential manner or it is expensive to collect the full information of training data. For example, feature selection in bioinformatics, where acquiring the entire set of features/ attributes for every training instance is exclusive due to the elevated cost in conducting wet lab experiments.

2. Existing System

Existing systems of online learning uses all the attributes or features in the training sets. Not always recommendable, because the real-time applications may use the datasets with high-dimensional instances or expensive to get all attributes or features. Many of them end at offline/batch learning, i.e., all features of the dataset is already present. But it is not applicable if the instances of training set are to be taken in live/online. There are three types Feature Selection algorithms supervised, Unsupervised and semi supervised. Selects features from labeled training set, select the important features by similarity of data or structures or partially labeled features.

Drawbacks:

- The complete training set must be available at once
- The training set must be available before the learning approach begins
- Uses all the attributes or features
- offline/batch learning
- Not recommended for using the datasets with high-dimensional instances or expensive to get all attributes or features

3. PROPOSED SYSTEM

We propose a solution of online feature selection (OFS) with a classifier using short and scalable set features. he training set is taken online and sequential. The goal is accurate prediction for an instance with the help of minimum number of active-features. The proposed system to improve the online batch learning using SVM based OML.

Advantages:

- Classifier uses short and scalable features.
- The training set is taken online and sequential.
- Accurate prediction
- OFS by learning with partial inputs.
- Uses only selected attributes or features
- Online learning
- Recommendable for using the datasets with high-dimensional instances or expensive to get all attributes or features.

Algorithm:

OML Algorithm: Machine Learning can use data mining techniques to develop models of algorithms for system performance or improve purpose.

BEGIN
1. Collect abstracts
2.Pre-processing
3. Text extraction
4. If 2c then
5. Informative
6. 3c-feature extraction

ISSN: 2278 – 1323

7.Else
8.Non-informative
9.goto preprocess step 3
10.End if
11.Update database
12.Calculate F-measure
13.Display R
14. END

models of algorithms for system performance or improve purpose.

4. Architecture Diagram:

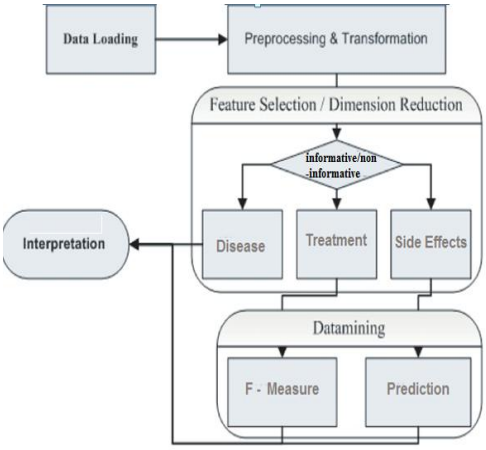

Fig 1: Architecture

Implementation

Modules

Here, the system is implemented by the following modules

1. OFS Learning
2. Online collection of Medical Abstracts
3. Information extraction(2c)
4. Relation Identification(3c)

Modules Description:

a) OFS Learning

The online feature selection approach with partial input information by employing a classical technique for making tradeoff between exploration and exploitation. In this approach, we will spend ε of trials for exploration by randomly choosing B attributes from all d attributes, and the remaining 1−ε trials on exploitation by choosing the B attributes for which classifier wt has non-zero values. Algorithm 4 shows the detailed steps of the proposed **OML** algorithm. Finally, Theorem 2 gives the mistake bound of algorithm.

b) Online collection of Medical Abstracts

The Medline is a health blog managed by the medical experts , it is a considered as a high knowledge base for online learning for Medical prediction systems . we collect the medical abstracts submitted by the medical experts in this blog, to extract the features to predict the disease - treatment relationships

c) Information extraction (2c)

The first task identifies sentences since Medline available abstracts that discuss concerning diseases and treatments. The task is like to a check of sentences restricted in the abstract of an object in order to near to the user only sentences that are known as containing important information (disease treatment information).It identifies whether sentences are informative, the data containing in order about disease and treatment, or not..

d) Relation Identification (3c)

This task has a deeper semantic measurement as well as it is alert on identifying disease treatment dealings in the sentences previously prefer as being informative.The focus is to by design classify which sentences include in sequence for the three semantic relations are Cure, Prevent, and Side Effect.

5. RESULTS

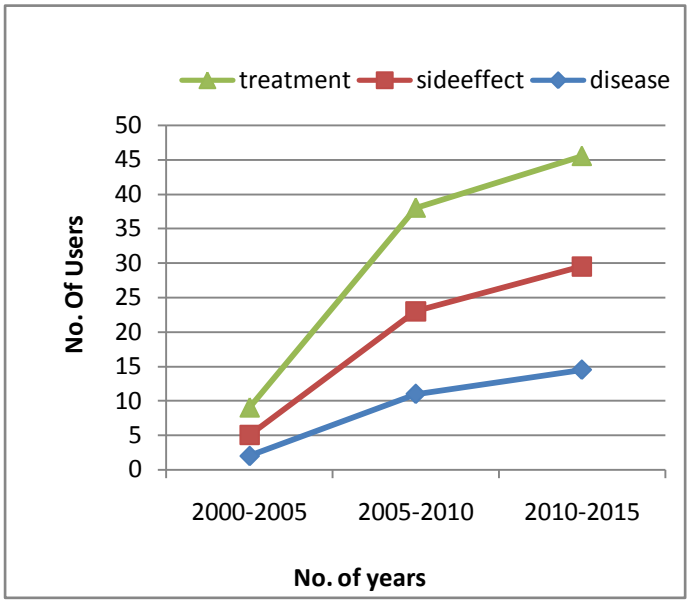

Fig 2: Performance analysis

Result analysis:

The above graph shows the result for calculating F-measure based on treatment and sideeffect.

6. CONCLUSION AND FUTURE WORK

In previously works data of any individual cannot access. It is obtained in terms of group which causes lots of space wastage and time consuming Process. We propose online learning methodology we are able to access any individual data are using with less duration of time and increase the availability of data. Future work may well longer our construction to other settings, for example, online multiclass categorization and deterioration problems, or to assist

ISSN: 2278 – 1323

undertake supplementary talented online learning tasks, such as online transfer learning or online AUC maximization.

7. References

[1] S.C.H. Hoi, J. Wang, and P. Zhao, "LIBOL: A Library for Online Learning Algorithms," Nan yang Technological Univ., 2012.

[2] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," technical report, Univ. of Toronto, 2009.

[3]J. Langford, L. Li, and T. Zhang, "Sparse Online Learning via Truncated Gradient," J. Machine Learning Research, vol. 10, pp. 777-801, 2009.

[4] H. Liu and L. Yu, "Toward Integrating Feature Selection Algorithms for Classification and Clustering," IEEE Trans. Knowledge and Data Eng., vol. 17, no. 4, pp. 491-502, Apr. 2005.

ISSN: 2278 – 1323