International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles available online http://www.ijoer.in

Vol.3., Issue.3, 2015

RESEARCH ARTICLE

ISSN: 2321-7758

TRAIT ASSORTMENT USING DISTRIBUTED SOLVING SET ALGORITHM

J.MAHESWARI[1], K.AMUTHA[2], R.KAYATHRI[3]

[1]PG Scholar, Department of CSE, Renganayagi Varatharaj College of Engineering, Sivakasi, India
[2]Assistant Professor, Department of CSE, Renganayagi Varatharaj College of Engineering, Sivakasi, India
[3]PG Scholar, Department of CSE, Renganayagi Varatharaj College of Engineering, Sivakasi, India

Article Received:12/05/2015 | Article Revised on:19/05/2015 | Article Accepted on:25/05/2015

ABSTRACT

Feature selection is very important process in data mining. Online learning represents a promising family of economical and scalable machine learning algorithms for large-scale applications. Most existing studies of online learning need accessing all the attribute/features of training instances. When information instances are of high dimensionality or it is exclusive to obtain the total set of attributes/features. Online feature selection aims, to select a small and fixed number of features for binary classification in an online learning. Modified perceptron and the sparse projection are used to classify the features in the existing system. The objective of the project is, to propose distributed method for detecting distance-based outliers in very large data sets. Our approach is based on the concept of outlier detection solving set, which is a small subset of the data set that can be also employed for predicting novel outliers. To compare their performance with three algorithms based on binary classifications To improve the classification accuracy by using distributed solving set.

Key Words— Feature Selection, Online Learning, Large-scale Data Mining, outlier detection, distributed algorithm

.

I. INTRODUCTION

Outlier detection is the data mining task whose goal is to isolate the observations which are considerably dissimilar from the remaining data . This task has practical applications in several domains such as fraud detection, intrusion detection, data cleaning, medical diagnosis, and many others. Unsupervised approaches to outlier detection are able to discriminate each datum as normal or exceptional when no training examples are available. Among the unsupervised approaches, distance-based methods distances to its nearest neighbors . These approaches differ in the way the distance measure is defined, but in general, given a data set of objects, an object can be associated with a weight or score, which is, intuitively, a function of its k nearest neighbors distances quantifying the dissimilarity of the object from its neighbors.

Data mining, also called knowledge discovery in databases. The process of discover interesting and useful patterns and relationships in

large volumes of data. It is widely used in business, science research and government security.

Feature selection is an important topic in data mining and machine learning. The objective of feature selection is to select a subset of relevant features for building effective prediction models. Most existing studies of feature selection are restricted to batch learning, which assumes the feature selection task is conducted in an off-line/batch learning fashion and all the features of training instances are given a priori. Online Feature Selection (OFS), aiming to resolve the feature selection problem in an online fashion by effectively exploring online learning techniques. Specifically, the goal of online feature selection is to develop online classifiers that involve only a small and fixed number of features for classification.

FS algorithms generally can be grouped into three categories: supervised, unsupervised, and semi-supervised FS. Supervised FS selects features according to labeled training data. Based on different selection criterions and methodologies, the existing supervised FS methods can be further divided into three groups: Filter methods, Wrapper methods, and embedded methods approaches. Filter methods choose important features by measuring the correlation between individual features and output class labels, without involving any learning algorithm; wrapper methods rely on a predetermined learning algorithm to decide a subset of important features. Although wrapper methods generally tend to outperform filter methods, they are usually more computationally expensive than the filter methods. Embedded methods aim to integrate the feature selection process into the model training process. They are usually faster than the wrapper methods and able to provide suitable feature subset for the learning algorithm. When there is no label information available, unsupervised feature selection attempts to select the important features which preserve the original data similarity or manifold structures.

Variable and feature selection have become the focus of much research in areas of application. These areas include text processing of internet documents and gene expression array analysis. The objective of variable selection is three-fold: improving the prediction performance of the predictors, providing faster and more cost-effective predictors and providing a better understanding of the underlying process that generated the data.

II.RELATED WORK

Group regularizes like group lasso have been extensively studied in both the statistics and machine learning fields. The objective of group lasso is to select a group of features simultaneously for a given task(s). The key assumption behind the group lasso regularizer is that if a few features in a group are important, then most of the features in the same group should also be important. However, in many real-world applications, we may come to the opposite observation. Consider the problem of multi-category document classification. The existing approaches for multi-task feature selection usually assume a positive correlation among the categories, namely, when one keyword is important for several categories, it is also expected to be important for the other categories. This positive correlation is usually captured by a group lasso regularizer, where a group is defined for every word w to include the feature weights of all categories for w. However, when our objective is to differentiate the related categories, we may expect a negative correlation among categories, namely, if word w is deemed to be important for one category, it becomes less likely for w to be an important word for the other categories. It is clear that such a negative correlation violates the assumption made by most of the existing approaches for multi-task feature selection. In order to capture the negative correlation among categories,we propose the exclusive lasso regularizer. Different from the group lasso regularizer, if one feature in a group is given a large weight, the exclusive lasso regularizer tends to assign small or even zero weights to the other features in the same group. Exclusive lasso regularizer is able to introduce competitions among variables and thus generate sparse solutions. This regularizer is applied to a multi-task feature selection setting and an efficient algorithm is given to solve the related optimization problem. Feature selection, which is known as a process of selecting relevant features and reducing dimensionality, has been playing an important role in both research and application. Feature selection can be conducted in a supervised or unsupervised manner, in terms of whether the label information is utilized to guide the selection of relevant features. Generally, supervised

feature selection methods require a large amount of labeled training data. It, however, could fail to identify the relevant features that are discriminative to different classes, provided the number of labeled samples is small. On the other hand, while unsupervised feature selection methods could work well with unlabeled training data, they ignore the label information and therefore are often unable to identify the discriminative features. Given the high cost in manually labeling data, and at the same time abundant unlabeled data is often easily accessible, it is desirable to develop feature selection methods that are capable of exploiting both labeled and unlabeled data. This motivates us to introduce semi supervised learning, into the feature selection process. Semi-supervised learning approaches can be roughly categorized into two major groups. The first group is based on the clustering assumption that most data examples, including both the labeled ones and the unlabeled ones, should be far away from the decision boundary of the target classes. The representative approaches in this category include tranductive support vector machine (SVM) and semi-supervised SVM. The second group is based on the manifold assumption that most data examples lie on a low-dimensional manifold in the input space. The method of semi-supervised SVM with manifold regularization has demonstrated good performance.

Variable selection refers to the problem of selecting input variables that are most predictive of a given outcome. Variable selection problems are found in many supervised and unsupervised machine learning tasks including classification, regression, time series prediction, clustering, etc. The objective of variable selection is two-fold: improving prediction performance and enhancing understanding of the underlying concepts in the induction model. Variable selection is a search problem, with each state in the search space specifying a subset of the possible attributes of the task. Genetic algorithms, population-based learning, and related Bayesian methods have been commonly used as search engines for the variable selection process. Particularly for SVMs, a variable selection method was introduced based on finding the variables that minimize bounds on the leave-one-out error for classification. The search of variable subsets can be efficiently performed by a gradient descent algorithm. Variable selection methods are often divided along two lines: filter and wrapper methods. The filter approach of selecting variables serves as a preprocessing step to the induction. The main disadvantage of the filter approach is that it totally ignores the effects of the selected variable subset on the performance of the induction algorithm. The wrapper method searches through the space of variable subsets using the estimated accuracy from an induction algorithm as the measure of “goodness” for a particular variable subset. Our approach (VS-SSVM) consists largely of two consecutive parts: variable selection and nonlinear induction. The selection of variables serves as a preprocessing step to the final kernel SVR induction. The variable selection itself is performed by wrapping around linear SVMs (no kernels) with sparse norm regularization. Such sparse linear SVMs are constructed to both identify variable subsets and assess their relevance in a computationally cheaper way compared with a direct wrap around nonlinear SVMs.

Online learning has been studied extensively in the machine learning. In general, for a misclassified example, most of the kernel based online learning algorithms will simply assign to it a fixed weight that remains unchanged during the whole learning process. Although such an approach is advantageous in computational efficiency, it has significant limitations. This is because when a new example is added to the pool of support vectors, the weights assigned to the existing support vectors may no longer be optimal, and should be updated to reflect the influence of the new support vector. We emphasize that although several online algorithms are proposed to update the example weights as the learning process proceeds, most of them are not designed to improve the classification accuracy. For instance, online learning algorithms are proposed to adjust the example weights in order to fit in the constraint on the number of support vectors; in example weights are adjusted to deal with the drifting concepts. It is designed to dynamically tune the weights of support vectors in order to improve the classification performance. In some trials of online learning, besides assigning a weight to the misclassified example, the proposed online learning algorithm also updates the weight for one of the existing support vectors. We refer to the proposed

approach as Double Updating Online Learning or DUOL for short. The key challenge in the proposed online learning approach is to decide which existing support vector should be selected for updating weight. An intuitive choice is to select the existing support vector that "conflicts" with the new misclassified example that is the existing support vector which on the one hand shares similar input pattern as the new example and on the other hand belongs to a class different from that of the new example. In order to quantitatively analyze the impact of updating the weight for such an existing support vector, we employ an analysis that is based on the work of online convex programming by incremental dual ascent.

III.SYSTEM DESIGN

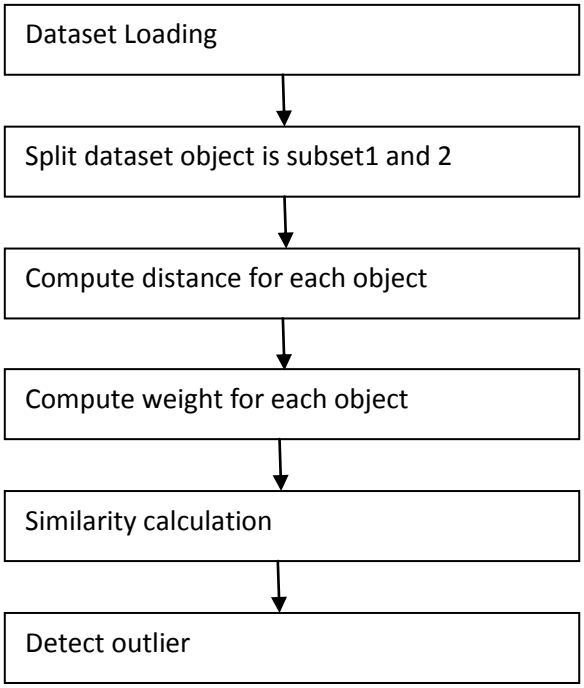

Fig1: System design

IV.PROPOSED WORK

Online Feature Selection (OFS), aiming to resolve the feature selection problem in an online fashion by effectively exploring online learning techniques. Specifically, the goal of online feature selection is to develop online classifiers that involve only a small and fixed number of features for classification. Feature selection is the technique of selecting subset of original features according to certain criteria. It is used for dimension reduction and hence, can be called as dimensionality reduction technique. Outlier detection solving set , which is a small subset of the data set that can be also employed for predicting novel outliers.

Modules

- **Subset of Dataset**
- **Estimate distance of each object**
- **Estimate weight of the each object**
- **Outlier detection**

The Distributed Solving Set algorithm is different, since it computes the true global model through iterations where only selected global data and all the local data are involved.

The core computation executed at each node consists in
the following steps:

1. Receiving the current solving set objects together with the current lower bound for the weight of the top nth outlier,
2. Comparing them with the local objects,
3. Extracting a new set of local candidate objects (the objects with the top weights, according to the current estimate) together with the list of local nearest neighbors with respect to the solving set and, finally,
4. Determining the number of local active objects, that is the objects having weight not smaller than the current lower bound.

A. Subset of dataset

First of all the user have to load the dataset. It contains number of times, glucose level, blood pressure, skin thickness, insulin level, body mass index, pedigree, age, class variables. This dataset is classified into two types. They are 0 and 1. 1 may be the positive for diabetes. And 0 may be the negative for diabetes. Based on the class label the data set are classified. It is called as the subsets of dataset. These subsets are stored into table called positive and negative.

B. Estimate distance of each object

For calculating the distance the average and standard deviation is used to find out. Average value is find out by summate the total values and divide it by the total instances. Based on the average value the standard deviation is calculated. The distance is calculated on the basis of standard deviation values. This distance is useful to find the weight of the objects.

C. **Estimate weight of each object**

The weight is calculated on the basis of the distance calculated from the dataset. After that the

clusters are formed. There are three clusters. On these three clusters the outliers are removed. The cluster formation is very important. For that the distributed solving set algorithm is used. Based on the algorithm the insulin level is divided into three categories. Based on the category of the insulin level the clusters are formed. The outliers are removed from these clusters.

D. Outlier detection

Outliers are needed to remove from the dataset. The outlier detection from the whole dataset will reduce the outliers in low level. But we formed three clusters. The outlier detection in these three clusters is highly accurate. The outliers are removed from the clusters on the basis of the weight values. The average is calculated for the remove the outliers. These outlier data are negative for diabetes. And the remaining data are positive for the diabetes. If the values are greater than the average means then it is considered as the positive diabetes. If the values are lower than the average means then it is considered as the negative diabetes.

V. EXPERIMENTAL RESULTS

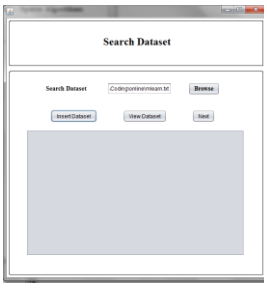

Fig2: Search Dataset

Fig3: Insert Dataset

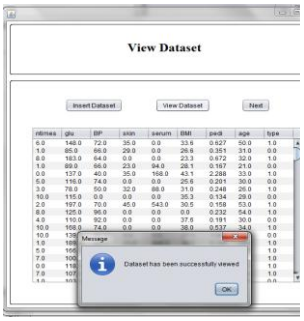

Fig4:View Dataset

Fig5:Subset Selection

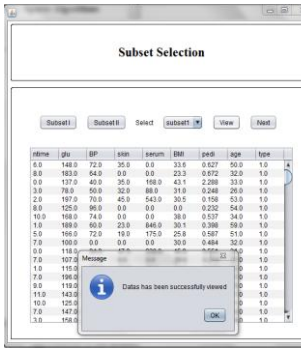

Fig6:Subset-I selection

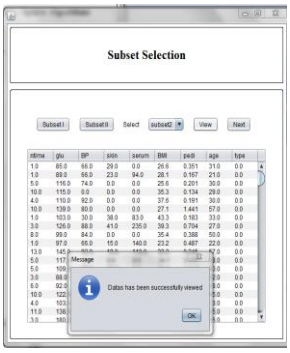

Fig7:Subset-II selection

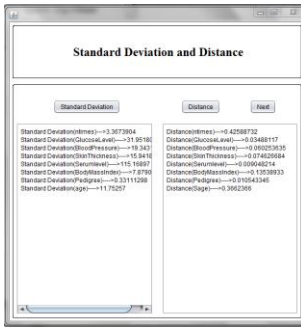

Fig7:Standard deviation and distance

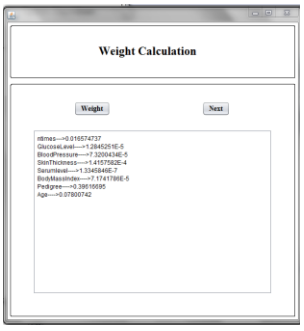

Fig8:Weight calculation

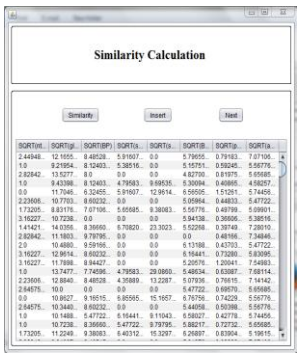

Fig9:Similarity calculation

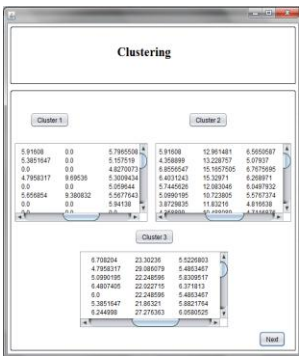

Fig10:Clustering

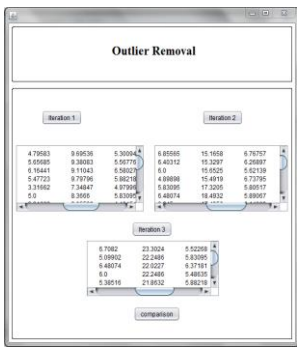

Fig11:Outlier removal

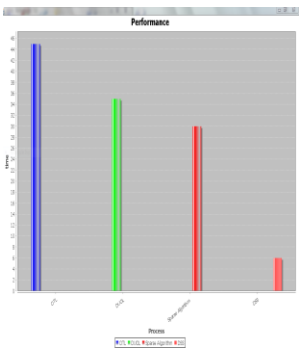

Fig12:Performace

VI.CONCLUSIONS AND FUTURE WORK

Feature selection is an important topic in data mining and machine learning, the objective of feature selection is to select a subset of relevant features for building effective prediction models. Online Feature Selection (OFS), which aims to select

a small and fixed number of features for binary classification in an online learning fashion. We presented the Distributed Solving Set algorithm, a distributed method for computing an outlier detection solving set and the top-n distance-based outliers.

VII. REFERENCES

[1] Y. Zhou, R. Jin, and S.C.H. Hoi, "Exclusive Lasso for Multi-Task Feature Selection", J.Machine Learning Research - Proc. Track, vol. 9, pp. 988-995, 2010.

[2] P. Zhao, S.C.H. Hoi, and R. Jin, "Double Updating Online Learning," J. Machine Learning Research, vol. 12, pp. 1587-1615, 2011.

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

[4] M. Unser Murray Eden, "Multi dimensionality reduction method using feature selection and feature extraction", International Journal of Artificial Intelligence & Applications (IJAIA), Vol.1, No.4, October 2010.

[5] F. Orabona, L. Jie, B. Caputo, "Multi Kernel Learning with Online-Batch Optimization", Journal of Machine Learning Research 13 227-253, 2012.

[6] J. Langford, Lihong Li, Tong Zhang, "Sparse Online Learning via Truncated Gradient", Journal of Machine Learning Research 10 (2009) 777-801.

[7] I. Guyon, A. Elisseeff, "An Introduction to Variable and Feature Selection", Journal of Machine Learning Research 3 (2003) 1157-1182.

[8] P. Bennett, Mark Embrechts, M. Breneman, "Dimensionality Reduction via Sparse Support Vector Machines", Journal of Machine Learning Research 3 (2003) 1229-1243.

[9] P. Zhao and S.C.H. Hoi, "OTL: A Framework of Online Transfer Learning," Proc. Int'l Conf. Machine Learning (ICML '10), pp. 1231-1238, 2010