Distributed Supervised Multiview Feature Selection for Efficient Big Data Analysis

[1]K. Mohana, [2]E. Ambika

[1]Assistant Professor, [2]M.Phil Scholar
[1,2]Department of Computer Science,
[1]Sri Ramalinga Sowdambigai College of Science and Commerce, Coimbatore, India
[2]Kovai Kalaimagal College of Arts and Science, Coimbatore, India.

Abstract: The rapid growth in popularity of economic activities can help to gather a huge amount of economic data. Though such data provides excellent opportunities for economic analysis, its poor quality, high-dimensionality and large-volume pose huge challenges on efficient analysis of economic big data. The traditional techniques offer an unsatisfactory performance while embracing the huge varieties of economic features. As a result, a new method was proposed for efficient analysis of high-dimensional economic big data based on innovative Distributed Feature Selection (DFS). Specifically, this method combines the economic feature selection and econometric model construction to reveal the hidden patterns for economic development. However, it requires the prior knowledge of the number of views during distributed feature selection process. Therefore, in this work, Distributed Supervised Multiview Feature Selection (DSMFS) method is proposed for big data analysis. Initially, data pre-processing is used to prepare the high-quality data. Then, an innovative distributed multiview feature selection is proposed to choose the significant and representative features from multidimensional dataset. In this technique, a group or view consists of homogeneous features, describing a unique data characteristic. Different views represent heterogeneous data characteristics. Consequently, a view can be represented by a few representative features in each view, and the information of heterogeneous views can be well kept by the remaining representative features. Finally, the experimental results demonstrate that the proposed DSMFS method can achieve higher performance for analyzing high-dimensional dataset than the existing DFS method.

***IndexTerms* – Big data, Feature selection, Supervised clustering, Homogeneity, Heterogeneity.**

I. INTRODUCTION

Big data is defined as the dataset whose dimension exceeds the capability of traditional dataset control systems in acquiring, storing, processing and analyzing. The challenge due to those 3V features i.e., volume, variety and velocity, has become the focus of learning techniques while dealing with big data. Also, redundancy and irrelativeness which are essential in big data with the aim of avoiding losing effective material, frequently create the mining process more essential. In data pre-processing phase, there are two approaches used such as data preparation and data reduction. Data preparation is required process which includes data cleaning, transformation, normalization, etc., for classification, clustering, etc. Data reduction is non-compulsory process which consists of feature selection, instance reduction, data compression, etc.

Learning from very large databases is a major issue for most of the current data mining and machine learning algorithms. This problem is commonly named with the term big data which refers to the difficulties and disadvantages of processing and analyzing huge amounts of data. It has attracted much attention in a great number of areas such as bioinformatics, medicine, marketing, or financial businesses, because of the enormous collections of raw data that are stored. Recent advances on Cloud Computing technologies allow for adapting standard data mining techniques in order to apply them successfully over massive amounts of data.

From high dimensional dataset, choosing less significant features is known as feature selection or feature reduction. It refers to the process of selecting few significant features/attributes from all features to reduce its feature size. Feature selection has facilitated data mining for its better performance of looking for correlated features from the actual dataset. It is one of the most significant data processing techniques and is often exploited to search correlated features and remove redundant or uncorrelated features from a feature set. Random or noisy features frequently disturb a classifier learning correct correlations and redundant or correlated features increase the complexity of a classifier without including any significant data to the classifier. A variety of feature selection techniques such as filter, wrapper and embedded approaches have been developed in the previous decades.

Typically, it is also known as variable selection, attribute selection or variable subset selection. It is a data mining technique aiming at choosing an optimal subset of features from the entire feature set that provides the best performance in terms of well-defined criteria. Here, a feature refers to an attribute of data which denotes the function of these data in a specific aspect. Since feature selection executes efficiently in simplifying the model, shortening the training time and reducing the variance of the model, researchers can understand the pattern of the data model more easily by using feature selection [1-2]. In big data processing systems, the major problem is scalability. The massive redundancy or irrelevance absolutely accounts for it, not only consuming computing resources; but also affecting processing performance. If this helpful data can be eliminated when valuable clues are retained, then the size of big data will be highly lowered and as a result, apart from the computation efficiency, the processing performance of big data will be increased. As a result, the feature selection for big data is the most significant as to obtain the feature subset with superior divisibility of considerable necessity.

To tackle this problem, DFS method [3] was proposed for efficient economic big data analysis. Initially, usability preprocessing, relative annual price computation, growth rate computation and normalization techniques were used to clean and transform the collected economic big data. Then, DFS method is proposed to choose the features related to economic development from high-dimensional economic data. After that, obtained most representative features were classified by using three-layer classification model. However, a prior knowledge is needed to select the features in multiple views that represent homogeneity and heterogeneous data characteristics. Also, it requires more effective solution to select the most significant features for big data analysis.

Hence in this article, DSMFS method is proposed for big data analysis. Initially, data pre-processing is used to prepare the high-quality data. Then, an innovative distributed multiview feature selection is proposed to choose the significant and representative features from multidimensional dataset. In this technique, a group or view consists of homogeneous features, describing a unique data characteristic. Different views represent heterogeneous data characteristics. Consequently, a view can be represented by a few representative features in each view, and the information of heterogeneous views can be well kept by the remaining representative features.

The rest of the article is prepared as follows: Section II presents the related works on feature selection for big data analysis. Section III explains the proposed methodology. Section IV illustrates the experimental results. Section V concludes the entire discussion.

II. Literature Survey

Zhang et al. [4] proposed a hybrid feature selection method combining Relief-F and mRMR for gene expression data. Relief-F is first used to look for a candidate gene set, and then the mRMR method is used to directly reduce redundancy for selecting a compact yet effective gene subset from the candidate set. Luo et al. [5] have proposed a two-step algorithm to combine the feature selectors for textual information in advertisements on the web. The algorithm first intersects two global feature selection results and then performs a local feature selection. Their experimental results indicate that their combination methods are efficient for a specific background. However, they cannot guarantee an optimal feature subset.

Wu et al. [6] defined relevance based on the exclusion of the conditional independence, whereas Kira and Rendell [7] described the RELIEF algorithm to estimate the weights of the features. Representative filter methods are RELIEF [7], FOCUS [8], and MIFS [9]. However, they have the following drawbacks: greatly relying on the stopping criteria (a threshold for determining when to stop these methods) and the mechanisms for calculating the importance of a feature. Besides, the strategy of seeking features is an influential factor on filter-based feature subset evaluation methods. Although the selection process of filters relies on the classification algorithms, the best filter measure is likely to be classifier specific, since different classifiers perform differently when combined with the same filter.

Xia et al. [10] proposed a feature ensemble plus sample selection method for domain adaptation in sentiment classification. This approach can yield significant improvements compared to individual feature ensemble or sample selection methods to take full account of two attributes, i.e. labeling adaptation and instance adaptation. In addition, some effective methods for feature selection problems have been proposed, such as improved Fisher score algorithm and enhanced Bare-bones Particle Swarm Optimization (BPSO). However, it concentrates only on search strategy of optimal subset to improve the performance.

III. Proposed Methodology

In this section, the proposed DSMFS method is explained in brief. This method has three phases such as data pre-processing, feature selection and classification. Initially, noise removal and missing value imputation are executed to increase the data usability. Also, the min-max normalization method is proposed for all features to avoid the influence of absolute values on the analytic results. Then, the pre-processed data are applied to the DSMFS method for choosing the most representative features/attributes. Further, the three-layer classification model is used for classifying the selected features.

The main objective of this work is to reduce the potentially large set of candidate attributes/features generated by the pre-process layer to a small set of potential attributes/features which are varied and similar to the attributes in the actual dataset. To achieve this problem, systematic attribute/feature selection method is proposed for big data analysis. The objectives of this method are to decide the important features by generalizing the distributed multiview subtractive clustering and predict the representative ones by designing the feature coordination-based clustering. Therefore, the complete utilization of the representative characteristics and their related important features are obtained to mine the direct and indirect effect on big data analysis.

3.1 Important Feature Selection

By approaching correlation analysis on big data, the important and representative records and indicators can be predicted. A Supervised Subtractive Clustering (SSC), i.e., a density-based distributed multiview clustering algorithm is a favorable method to investigate the correlations between data samples. It assumes that each data point is a potential cluster center and calculates a measure of the likelihood based on the density of surrounding data points and multi-views. In this way, it can construct the relationships among all the data points and views. When decomposing the relationships to a same attribute, the contribution of the attribute to preserve the relationships can be achieved. According to this idea, the SSC is used to identify the important indicators for economic analysis.

The considered dataset has decades of records with a range of indicators sorted by year. For each record, its density value contributed by other records can be calculated as follows:

$$D_i = \sum_{j=1}^{n} e^{\left[-\frac{\left\|x_i - x_j\right\|^2}{(0.5r^*)^2}\right]} \qquad (1)$$

In Eq. (1), the dataset $\{x_1, x_2, \dots, x_n | y_i, i = 1, \dots, n\}$ is n data pairs with $x_i \in \mathbb{R}^m$ where m refers to the dimension number, n refers to the object number, y_i refers to the class label of object $x_i, y_i \in \{1,2, \dots, c\}$ and c denotes the class number, the objects are written as $X = [x_1, x_2, \dots, x_n]^T \subset \mathbb{R}^{n\times m}$ and the corresponding label vector is as $Y = [y_1, y_2, \dots, y_n]^T \subset \mathbb{R}^n$. The m feature vectors are denoted as $F = [f_1, f_2, \dots, f_n]$ and i^{th} feature vector $f_i \subset \mathbb{R}^{m\times 1}$.

3.2 View Generation by Affinity Propagation

View generation denotes decomposition of the original data set into multiple disjoint groups, each of which can be seen as a view. Here, Affinity Propagation (AP) is used for view generation. The similarities $S(i,j)$ between data vectors x_i and x_j are obtained to indicate how well j^{th} data vector x_j is suited to be the exemplar of x_i as:

$$S(i,j) = -\left\|x_i - x_j\right\|^2 \qquad (2)$$

The self-similarity or preference is set as:

$$S(t,t) = \frac{\sum_{i,j=1,i\neq j}^{n} S(i,j)}{n\times(n-1)}, 1 \leq t \leq n \tag{3}$$

The data x_i transmits responsibility $r(i,j)$ to x_j. $r(i,j)$ reflects the accumulated evidences for how well x_j serves as the exemplar of x_i. The availability $a(i,j)$ is transmitted from x_j to x_i for reflecting the accumulated evidence of how suitably selects x_j as its exemplar. The update equations can be written as follows:

$$r(i,j) = s(i,j) - \max_{j'\neq j}\{a(i,j') + s(i,j')\} \tag{4}$$

$$a(i,j) \leftarrow \begin{cases} \min\{0, r(j,j) + \sum_{i'\neq i,j} \max\{0, r(i',j)\}\}, & i \neq j \\ \sum_{i'\neq i} max\{0, r(i',j)\}, & i = j \end{cases} \tag{5}$$

Then, the exemplars are recognized as follows:

$$\mathcal{V} = \arg\max_{k}\{r(i,k) + a(i,k)\} \tag{6}$$

Finally, L feature groups or views (V) are generated and l^{th} feature group is written as F_l. The views are heterogeneous and each view is composed of homogeneous features. This step is critically important to sufficiently represent data. Based on this, a high density value corresponds to a data sample with many neighborhood data samples and views. Hence, the data sample with the highest density is chosen as the initial cluster center.

After that, the samples close to the initial cluster center being chosen as the other centers of clusters is avoided by an amount of density proportional is subtracted from each sample to its distance from the first cluster center. After the reduction, the data sample with the highest remaining density is selected as the second cluster center and the density of each data sample is further reduced according to its distance to the second cluster center. In general, after k^{th} cluster center x_{c_k} is obtained with density D_{c_k}, the density of each data sample is updated by:

$$D_{i,\mathcal{V}} = D_{i,\mathcal{V}} - D_{c_k} e^{\left[-\frac{\|x_i - x_j\|^2}{(0.5r^*)^2}\right]} \tag{7}$$

The processes of finding new cluster centers and reducing the density for each data sample iterate until the remaining densities of all data samples are bounded by some fraction of the density D_{c_1} of the first cluster center. The termination criterion is ${D_{c_k}}/{D_{c_1}} > \epsilon$.

In Eq. (7), the density value $D_{i,\mathcal{V}}$ is affected by the sample feature value in V i.e., $A_{a,V}$. Therefore, the contribution of attribute a to sample i and view V as:

$$I(i,V,a)_k = \sum_{p=1}^{n} \left|\frac{\partial D_{i,\mathcal{V}}}{\partial x_{pa}}\right| \tag{8}$$

This refers to how much data there is to cluster the i^{th} sample and V^{th} view with a^{th} feature. According to Eq. (8), $\frac{\partial D_{i,\mathcal{V}}}{\partial x_{pa}}$ has the following four forms in which part of the records are chosen for computing the feature contribution as:

(i) When $i = p, j = p$

$$\frac{\partial D_{i,\mathcal{V}}}{\partial x_{pa}} = \frac{4\left(x_{pa} - x_{c_k a}\right)}{(r^*/2)^2} e^{\left(\frac{-2\sum_{r=1}^{m}\left(x_{pr} - c_{c_k r}\right)^2}{(r^*/2)^2}\right)} \tag{9}$$

(ii) When $i = p, j \neq p$

$$\frac{\partial D_{i,\mathcal{V}}}{\partial x_{pa}} = \left(\frac{2(x_{pa} - x_{ja})}{(r^*/2)^2} e^{\left(\frac{-\sum_{r=1}^{m}\left(x_{ir} - x_{jr}\right)^2}{(r^*/2)^2}\right)} + \frac{2\left(x_{pa} - x_{c_k a}\right)}{(r^*/2)^2} e^{\left(\frac{-\sum_{r=1}^{m}\left(x_{pr} - x_{c_k r}\right)^2}{(r^*/2)^2}\right)} e^{\left(\frac{-\sum_{r=1}^{m}\left(x_{c_k r} - x_{jr}\right)^2}{(r^*/2)^2}\right)}\right) \tag{10}$$

(iii) When $i \neq p, j = p$

$$\frac{\partial D_{i,\mathcal{V}}}{\partial x_{pa}} = \left(\frac{2(x_{ia} - x_{pa})}{(r^*/2)^2} e^{\left(\frac{-\sum_{r=1}^{m}\left(x_{ir} - x_{pr}\right)^2}{(r^*/2)^2}\right)} - \frac{2\left(x_{c_k a} - x_{pa}\right)}{(r^*/2)^2} e^{\left(\frac{-\sum_{r=1}^{m}\left(x_{c_k r} - x_{pr}\right)^2}{(r^*/2)^2}\right)} e^{\left(\frac{-\sum_{r=1}^{m}\left(x_{ir} - x_{c_k r}\right)^2}{(r^*/2)^2}\right)}\right) \tag{11}$$

(iv) When $i \neq p, j \neq p, \frac{\partial D_{i,\mathcal{V}}}{\partial x_{pa}} = 0$

Here, c_k is the k^{th} cluster center selected by SSC. Therefore, the significance of a^{th} feature for selecting k^{th} representative record is defined as:

$$I(a)_k = \sum_{i=1}^{n} I(i,V,a)_k \tag{12}$$

The features with the higher-ranking value contain more information of clusters than others. Thus, the most important features are obtained and given to the classifier for analyzing the big data effectively.

Algorithm: Important Feature Selection

Input: Data matrix $X \in \mathbb{R}^{n\times m}$ and parameter ϵ, σ

Output: Important features and cluster centers for X

```
    Initialize the neighbourhood radius $r^* = \sqrt{\sum_{j=1}^{n}\sum_{i=1}^{n} \|x_i - x_j\|^2 / n(n-1)}$ and Euclidean distance matrix $G$ between data samples;
    $\boldsymbol{for}(each\ data\ sample\ x_i \in X, i = 1, \dots, n)$
        Compute the density $D_i$ as Eq. (1);
        Compute the similarity $S(i,j)$ as Eq. (2);
        Compute responsibility $r(i,j)$ of $x_i$ as Eq. (4) and availability as Eq. (5);
        Obtain the $L$ feature groups or views $(V)$;
    end for
    The sample with the highest density $D_{c_1}$ is chosen as the primary center. Set $k = 1$;
    while $\left({D_{c_k}}/{D_{c_1}} > \epsilon\right)$
        for(each sample $x_i \in X, i = 1, \dots, n - k$)
            Update the density $D_{i,V}$ as Eq. (7);
        end for
        The new center with the highest density $D_{c_k}$ is chosen. Set $k = k + 1$;
        for(attribute/feature $a \in F$, $F$ is a feature set of $X$)
            for(each data sample $x_i \in X, i = 1, \dots, n$)
                Compute the contribution of attribute $a$ to sample $i$ and view $V$ in clustering as Eq. (8);
            end for
            Sum the contributions of attribute $a$ to all samples and views as Eq. (12);     //The significance of $a^{th}$ feature for selecting $k^{th}$ representative record
        end for
    end while
    The features with $I = {\sum_{j=2}^{k} I(a)_j}/{(k-1)} > \sigma$ are chosen;     //$I$ is the significance of $a^{th}$ feature for clustering
```

IV. Result and Discussion

In this section, the performance of proposed DSMFS method is evaluated and compared with existing DFS method by using Java. In this experiment, the "Adult Income Dataset" is used which is available in the UCI machine learning repository. This dataset consists of 48842 records and a binomial label that indicates the salary of 50K USD. Also, 76% of the records have a class label of <50K. The dataset is split into training and testing set. The training dataset has 32561 records and the testing dataset has 16281 records. There are 14 attributes consisting of eight categorical and six continuous attributes such as age, final weight, education number, capital gain, capital loss and hours per week.

The employment class describes the type of employer such as self-employed or federal and occupation describes the employment type such as farming, clerical or managerial. Education contains the highest level of education attained such as high school or doctorate. The relationship attribute has categories such as unmarried or husband and marital status has categories such as married or separated. The other nominal attributes are country of residence, gender and race. The comparison is made in terms of precision, recall and accuracy.

4.1 Precision

It is defined as the value which is evaluated for feature selection at True Positive (TP) prediction and False Positive (FP) prediction. It is measured as follows:

$$Precision = \frac{TP}{TP+FP} \quad (13)$$

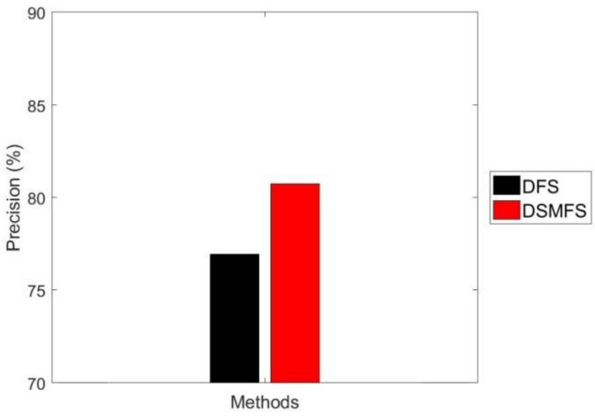

Fig.1 Comparison of Precision

Fig. 1 shows the comparison of precision for DFS and DSMFS methods. From this analysis, it is concluded that the DSMFS method achieves 4.93% higher precision than the DFS method.

4.2 Recall

It is defined as the value which is evaluated at TP and False Negative (FN) rates. It is measured as follows:

$$Recall = \frac{TP}{TP+FN} \quad (14)$$

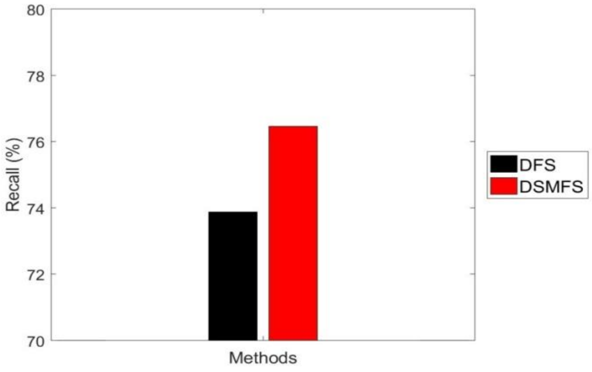

Fig.2 Comparison of Recall

Fig. 2 shows the comparison of recall for DFS and DSMFS methods. From this analysis, it is concluded that the DSMFS method achieves 3.48% higher recall than the DFS method.

4.3 Accuracy

It is the proportion of true results (both TP and TN) among the total number of cases examined.

$$Accuracy = \frac{TP+TN}{TN+TP+False\ Positive\ (FP)+ FN} \qquad (15)$$

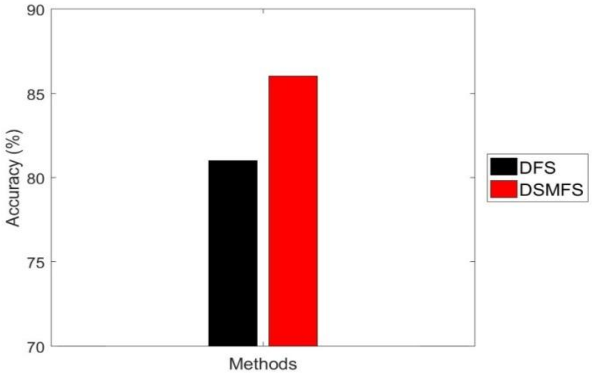

Fig.3 Comparison of Accuracy

Fig. 3 shows the comparison of accuracy for DFS and DSMFS methods. From this analysis, it is concluded that the DSMFS method achieves 6.2% higher accuracy than the DFS method.

V. Conclusion

In this paper, DSMFS method is proposed for improving the performance of big data analysis. In this method, novel feature selection method is proposed to learn the important features from the high-dimensionality, huge-volume, and low-quality economic data for economic model construction. Initially, data preprocessing and normalization techniques are applied to reduce the noise and transform the collected big data. After that, a DSMFS is proposed to construct a feature selection model which selects the important features and identifies the representative ones of big data in horizontally and vertically. With the representative attributes extracted by the feature selection model, a collaborative model is constructed for analyzing the big data. Finally, the experimental outcomes demonstrated that the proposed feature selection method has better performance than the existing feature selection method for big data analysis.

References

[1] Rong, M., Gong, D., & Gao, X. (2019). Feature Selection and Its Use in Big Data: Challenges, Methods, and Trends. IEEE Access, 7: 19709-19725.

[2] Li, J., & Liu, H. (2017). Challenges of feature selection for big data analytics. IEEE Intelligent Systems, 32(2): 9-15.

[3] Zhao, L., Chen, Z., Hu, Y., Min, G., & Jiang, Z. (2016). Distributed feature selection for efficient economic big data analysis. IEEE Transactions on Big Data, 4(2): 164-176.

[4] Zhang, Y., Ding, C., & Li, T. (2008). Gene selection algorithm by combining reliefF and mRMR. BMC genomics, 9(2): S27.

[5] Luo, Y., Li, Y., Zhou, C., & Xu, C. (2011). Combining feature selectors in a product advertisement classification system. In The First Asian Conference on Pattern Recognition, pp. 184-188. IEEE.

[6] Wu, X., Yu, K., Ding, W., Wang, H., & Zhu, X. (2012). Online feature selection with streaming features. IEEE transactions on pattern analysis and machine intelligence, 35(5): 1178-1192.

[7] Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In Machine Learning Proceedings 1992, pp. 249-256. Morgan Kaufmann.

[8] Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. IEEE Transactions on neural networks, 5(4): 537-550.

[9] Almuallim, H., & Dietterich, T. G. (1994). Learning boolean concepts in the presence of many irrelevant features. Artificial Intelligence, 69(1-2): 279-305.

[10] Xia, R., Zong, C., Hu, X., & Cambria, E. (2013). Feature ensemble plus sample selection: domain adaptation for sentiment classification. IEEE Intelligent Systems, 28(3): 10-18.