ISSN:0975-9646

Survey on Semi-Supervised Feature Selection in Datamining

Kalaiselvi.R[#1], Premadevi.P ME[*2], Hamsathvani.M[#3]

[#] *PG scholar Dept of CSE, Angel College Of Engineering And Technology,*
Tiruppur, Tamilnadu, India
[*]*Assistant Professor, Angel College Of Engineering And Technology,*
Tiruppur, Tamilnadu, India

Abstracts: Data mining is one of the knowledge extraction process used to discover the knowledge from large datasets and convert it into useful information. Various techniques implemented for this extracting process. In machine learning, feature selection is most important one to extract the feature from large dataset. Feature selection involves subset generation, assessment and terminating technique for achieving relevant data. This survey shows various machine learning techniques for extracting relevant data. Extraction is not an easiest one, they where many problems occurred due to selection of irrelevant data and main problem in accuracy of the data. This survey shows the problems and challenges of machine learning techniques for achieving the better accuracy from supervised, unsupervised, and semi-supervised learning techniques.

Keywords: Data mining, Machine learning, feature selection, relevant data, and accuracy.

I. INTRODUCTION:

Even with today's advanced computer technologies (e.g., machine learning and data mining systems) extracting data is critical one where data set are widely used in machine learning and data mining (DM) tasks. In Data Mining, discovering knowledge from data can still be fiendishly hard due to the characteristics of the computer generated data. In large dataset data are represented in feature value. Size of dataset may be measured in no. of features and no. of instance. In machine learning, **Feature selection** (FS) is one of the useful for data analyzing process; it proves which features are significant for prediction, and how these features are associated [1]. FS involves subset generation, subset assessment, terminating the search while achieve the redundant one. While in feature selection not all the features where relevant. Feature selection is the effective means to identify relevant features for decreasing no. of dimensionality [2]. Feature selections are labeled as (i) *Relevant:* A relevant feature is one which is related to the minimum cardinality for achieving the high predictive data. (ii) Irrelevant: Irrelevant features does not having any control on the output here the values are generated at random level for each data. (iii) *Redundant:* unwanted features occurred in the data. Selecting an original subset feature to the relevant one is not an easy one. Under *supervised learning*, feature selection is one of the popular one. Feature selection in supervised learning exploit predictive accuracy for some function where supervised learning is based on training data. Training data is a model data which we already assigned to the data source with a right classification and regression. In supervised learning it maps the set of *input* variables *X* and an *output* variable *Y* to predict the outputs for unlabeled data. In *supervised feature selection*, if we give a small size of labeled data means they didn't provide sufficient information to the target one; hence it may remove many relevant features. Classification techniques are widely used for supervised learning for the labeled data. In *unsupervised feature selection,* it uses a large no. of unlabeled data (i.e) learning without training data, hence this ignores label information and it generates the downgrade performance. Clustering techniques are mostly used for unsupervised learning to find out groupings of similar objects in data. By providing labeled and unlabeled didn't provide an expected outcomes, so the missing of labeled and unlabeled provide an expected outcomes, which is called semi-supervised learning. Semi-supervised learning (SSL) is a class of machine learning techniques. The clustering is the important one for most semi-supervised learning algorithms. It is estimate by its suitability with both labeled and unlabeled data by using a relevant feature. *Semi-supervised feature selection* is used to find the most informative pattern subsets [3]. *Semi-supervised feature selection* is the combination of Classification and Clustering techniques. Not all the algorithms give the accuracy for feature selection. This survey tells the various existing techniques for data extraction in the machine learning techniques using feature selection mechanism.

II. LITERATURE REVIEW

Feature selection is a vital topic in data mining generally for high dimensional datasets. In today environments, they where large amount of data are there to extracting the data from dataset is one of the critical one. For the extraction process feature selection helps to extract the relevant feature or data from the large no. of dataset. Feature selection is one of the best tools; it aims to reduce dimensionality for building comprehensible learning models with good generalization performance. The goal of this survey is to provide a comprehensive review of different feature selection of learning technique of various researchers and its varying.

Jennifer G. Dy and Carla E. Brodley studied the problem of automated feature subset selections for unlabeled data [4]. In the presents of irrelevant data it may be an over fitted. Feature selection is used to identify the relevant data by using some matrix like measurements, distance, vectors, attributes, etc… they were many problem occurred during feature selection. The problems were, during feature selection they need to find the number of clusters in combinations of the data. And have to perform the normalizing the feature selection with relevant data by using the dimension process.

Jennifer G. Dy and Carla E. Brodley proposed the wrapper framework using FSSEM (feature subset selection using EM clustering) [4]; by using this method, the search mechanism, feature selection method and feature normalization can be easily tackled for cluster method. EM which is referred to Expectation-Maximization is used to estimate the maximum possibility constraint of a finite Gaussian mixture. Although we examined the wrapper framework using FSSEM, the search method, feature selection criteria (especially the *trace* criterion), and the feature normalization scheme can be easily applied to any clustering method and maximize the predictive accuracy. This technique solved the above issues. For the normalization problem here they use cross-projection normalization scheme to eliminate the bias. For evaluating the feature subset under the above method, here they use biases of ML (Maximum Likelihood) and scatter separability with respect to measurement of data by using filter method. Their results address the problem and give better result for extraction the data.

D Zhang el at studied the problem of high dimensionality data in supervised and unsupervised feature selection. Using FLD (Fisher Linear Discriminant) for supervised approach does not provide better results on labeled data, where as using a PCA (Principal Component Analysis) in unsupervised also trying to preserve the global data but not extract preservation done hence unsupervised approach is an unlabeled data [5].

For the above problem D Zhang el at proposed a SSDR (Semi-Supervised Dimensionality Reduction) algorithm. In Semi-Supervised learning, it learns from both a labeled and unlabeled data. Unlabeled data are gladly available and the labeled data are quite expensive. Exploiting data is an important issue in DM. Here we use data exploiting in two form of constrains i.e. the same class (*must-link* constraints) or different classes (*Cannot-link* constraints) together with unlabeled data. By using SSDR, it shows low-dimensional space [5]. SSDR leads to considerable improvements in embedding, classification and clustering over conventional dimensionality reduction methods.

Zenglin Xu et al studied the problem of concave-convex optimization method; here the small amount of trained data is usually inadequate for recognizing the relevant features, it is the main issues for feature selections and problem arising from semi-supervised feature extraction, i.e. extracting the data from unlabeled data, hence semi-supervised feature extraction is the combination of labeled and unlabeled data [6].

Zenglin Xu et al proposed a novel discriminative semi-supervised feature method based on the maximum margin principle and the manifold regularization. This method is used to select the features by maximizing the margin between the different labels and distributes the generated data. Here the researchers used the embedded feature selection method whereas feature selection may be any one of these filter, wrappers and embedded method. With the use of embedded method it able to find more discriminate features. This proposed system is used to solve a concave-convex optimization problem. It solves the problem with the limited no. of dataset only but it didn't concentrate on big datasets [6].

Ianisse Quinzán studied the problem of extracting the data with a limited no. of labeled sets [7]. Where in unlabeled data extraction is improved accuracy because no training data may provide, but in labeled data they extract the feature based on the trained data so loss of data may occur because of eliminating the real data from the datasets. To find an optimal subset of features is risky factor under labeled data.

Ianisse Quinzán et al Proposed a technique based on Conditional Mutual Information and Conditional Entropy [7]. Here the researchers present a filter method for feature selection. For this filter method of feature selection a new techniques were built called hybrid method of semi-supervised which the combination of supervised and unsupervised (i.e. combination of labeled and unlabeled data). By using this approach it utilize a dissimilarity measure between each pair of features. It uses the conditional Mutual to give the preferences to the mutual one.

III. ACCURACY OF FEATURE SELECTIONS IN MACHINE LEARNING

Feature Selection is one of the most important tasks for various activities like machine learning, pattern recognition, data extraction, web service extraction, text mining, image processing, etc. The task of selection the best feature from the noisy feature is known as feature selection, variable selection or subset selection. The fundamental of feature selection is to avoid the noisy feature (data) and get the accurate data. Accuracy calculation is one of the meaningful and challenging factors under feature selection. Feature selection is most commonly used in machine learning process. Here the most machine learning techniques where supervised learning, unsupervised learning and semi-supervised learning used for sub-set feature selection [2, 8].

a. Feature selection in supervised learning

Supervised Learning (SL) is one of the classification methods. It is based on the labeled data (trained data). The labeled data are the sample data from original data source or data sets with correct classification which we already assigned. This solves the diverse problems. Fisher Linear Discriminant (FLD) is an example of supervised feature extraction process [base paper]. It can estimate the value by co-relations of the class label. It extracts the feature with the best discriminant ability of trained data. It estimates the importance of features in the bases of how well their values distinguish between the instances of the same and different classes that are near to each other [9, 10]. It is easy to extract the relevant data from the near data by two steps. (i) Training step: This process is used to learn the data to extraction process from the trained data. (ii) Prediction step: This process is used to assign the value to the test data and predict data for hidden data. They were many drawbacks under SL, the first one is difficult to collect the supervision or labels, it is expensive one for huge data, the main drawback is because of labeled data they may problem occur due to eliminating the relevant data in the trained data, because some features are likely irrelevant to each other. SL can predict and assume the testing samples of trained data only but failed in untrained data i.e. unlabeled data, hence the accuracy matrix is very low from this labeled data.

b. Feature selection in unsupervised learning

Unsupervised Learning (UnSL) is one of the clustering methods. It is used to identify hidden patterns in unlabelled data (untrained data) [9]. It is one of the learning algorithms, used to learn and organize the data without any knowledge of reference data. It extracts the data automatically [10]. Clustering is a common unsupervised learning method used to identify group structure in set of data. This is most commonly used for untrained dataset to extract the relevant information by checking constrains from nearest one, whereas UnSL algorithms simply give the output from the incoming data and no training process is needed [11]. It involves long learning time for extracting the feature. Then accuracy matrix is limited due to lack of pertinent with the dataset and it cannot provide adequately accurate results for the successive dataset.

c. Feature selection in semi-supervised learning

Semi supervised is one of the recent trends in machine learning techniques. Semi-supervised Learning (SSL) is one of the hybrid methods, which is the combination of Supervised and Unsupervised Learning [12]. It learns the data from the combination of both labeled and unlabeled data and evaluated by its fitness with both labeled and unlabeled data. Extracting the data using three stages *(i) Constraint Selection:* it is one of the primary method, used to specify if the two instance in the same class is consider to be a *must-link class* and if they were different class then it considered to be a *cannot-link class* [13]. These constraints have almost proven to have very positive effects on the learning performance. *(ii) Feature Relevance:* it is used to select the features with the better respect of data structure. Choosing a relevant feature is to minimize the dimension problem. Then final one is *(iii) Redundancy Analysis:* it is used to analyze the unnecessary data and then eliminate it to provide more accuracy for our method. It consists of many steps for redundancy analysis, selecting the irrelevant feature by ranking algorithm, construct map between irrelevant features, find the corresponding irrelevant features then finally eliminate the entire irrelevant feature [14]. By doing this steps we get an accurate data from this semi-supervised learning. It give better improved performance for several datasets most commonly in real world datasets, such as person identification in camera clips through image processing, GPS-based map refinement [15] and landscape detection from hyper spectral data, etc. it exploits the knowledge of the irrelevant distributions from the unlabeled samples and also it utilizes the label information offered by the labeled samples [16]. Most researchers concentrate on Semi supervised learning method, it give more accuracy when compared to the existing two learning methods.

IV OUTCOMES OF SURVEY

- With the increasing no. of computer technology data storage also increased. Extracting the relevant data is not an easiest one. Extraction process may also be done through feature selection mechanism.
- Feature selection is one of the most vital tools in data mining to extract the data from the datasets.
- Extracting the data is important one to choose the relevant data from datasets.
- We have studied and analyze the problem of extracting data through learning techniques of various researchers' approaches. Hence it does not solve the accuracy problem.
- Because of noisy data in the datasets it didn't provide better accuracy, by removal of noisy data then only we get a better accuracy.
- Lot of techniques was implemented for this removal of noisy data. Feature selection is one of the machine learning techniques to remove noisy data. There are lot of approaches for the extracting the data.
- In supervised learning in consists of labeled data, so it select the trained data only and it eliminate most data from the dataset and also labeled data are expensive one.
- In Unsupervised Learning, it is based on unlabeled data, so it can select the relevant data in accurate manner with lot of time taken process.
- In semi-supervised method, it consists of labeled and unlabeled one, hence process is studies how to better identify the relevant features with separation to other class data and effectively exploring the data from both huge and low level of labeled and unlabeled data

Table 1: Comparison table for feature selection under machine learning

Machine Learning	**Accuracy of feature selections**
Supervised feature selection	Eliminate relevant data
Unsupervised feature selection	Downgrade performance, high computational time.
Semi-Supervised feature selection	Increasing no. of accuracy, dimension reduction and improved performance.

V. CONCLUSION

In this survey, we study a comprehensive overview of various algorithms of feature selection in machine learning techniques. This survey challenges the problem of accuracy matrix and improving the accuracy matrix. With the tremendous growth of dataset extraction is most difficult one. Extracting data is one of the feature selection methods. For extracting process here we show a lot of existing approaches and they were problem occurred due to higher time complexity, difficult to scale high dimensionality, not occurred data are extracted and it is expensive for huge data, from various learning approaches like supervised and unsupervised method. To solve the above problem this survey presents the combination of supervised and unsupervised method i.e. semi-supervised paradigms. By using this semi-supervised method, it gives a better accuracy compared to existing techniques.

REFERENCES

[1] Ladha, L., and T. Deepa. "Feature selection methods and algorithms."*International Journal on Computer Science and Engineering* 3.5 (2011): 1787-1797.

[2] Kim, Yong, W. Nick Street, and Filippo Menczer. "Feature selection in data mining." *Data mining: opportunities and challenges* 3.9 (2003): 80-105.

[3] Zheng Zhao and Huan Liu, "Semi-supervised Feature Selection via Spectral Analysis", SDM 2007.

[4] J. G. Dy and C. E. Brodley, "Feature selection for unsupervised learning," *J. Mach. Learn. Res.*, vol. 5, pp. 845–889, Aug. 2004.

[5] Daoqiang Zhang, Zhi-Hua Zhou and Songcan Chen, "Semi-Supervised Dimensionality Reduction".

[6] Zenglin Xu, Irwin Kin, Michael Rung-Tsong Lyu and Rong Jin, "Discriminative Semi-Supervised Feature Selection Via Manifold Regularization", IEEE trans 2010.

[7] Ianisse Quinzán, José M. Sotoca, Filiberto Pla, "Clustering-based Feature Selection in Semi-supervised Problems" 2009.

[8] Quinzán, Ianisse, José Martínez Sotoca, and Filiberto Pla. "Clustering-Based Feature Selection in Semi-supervised Problems." *ISDA*. 2009.

[9] Sathya, R., and Annamma Abraham. "Comparison of supervised and unsupervised learning algorithms for pattern classification." *Int J Adv Res Artificial Intell* 2.2 (2013): 34-38.

[10] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in *Proc. 24th Int. Conf. Mach. Learn.*, Corvallis, OR, USA, 2007.

[11] Cai, Deng, Chiyuan Zhang, and Xiaofei He. "Unsupervised feature selection for multi-cluster data." *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010.

[12] Neethu Innocent 1, Mathew Kurian, "Survey on semi supervised classification methods and Feature selection". 2013.

[13] M. Hindawi, K. Allab, and K. Benabdeslem, "Constraint selection based semi-supervised feature selection," in *Proc. IEEE ICDM*, Vancouver, BC, Canada, 2011, pp. 1080–1085.

[14] V. Jothi Prakash and Dr. L.M. Nithya, "A Survey On Semi-Supervised Learning Techniques". IJCTT– Feb 2014.

[15] Y. Cheng, and Y. Cai, "Semi-Supervised Feature Selection Under Logistic I-RELIEF Framework", 19th International Conference on Pattern Recognition, ICPR 2008, IEEE, Tampa, Florida, USA, December 2008, pp 1-4.

[16] Barkia, Hasna, Haytham Elghazel, and Alex Aussem. "Semi-supervised feature importance evaluation with ensemble learning." *Data Mining (ICDM), 2011 IEEE 11th International Conference on*. IEEE, 2011.