International Journal of Soft Computing 9 (5): 303-307, 2014
ISSN: 1816-9503

Improving Performanc of Supervised Learners Using Unsupervised Variable Selection Algorithm: A Novel Approach

[1]D. Asir Antony Gnana Singh, [3]S. Appavu Alias Balamurugan and [2]E. Jebamalar Leavline
[1]Department of CSE, [2]Department of ECE, Bharathidasan Institute of Technology,
Anna University, Tiruchirappalli, India
[3]K.L.N. College of Information Technology, Sivagangai, India

Abstract: Prediction influences the technological advancement in various sectors includes in finance for predicting the behavior of the stock market, in sports for predicting the outcome of the event, in opinion polls for predicting the outcome of the election and in many applications predicting their related unseen or unknown data. The prediction can be performed by the supervised learners also known as classifiers and their performance mainly relay on the variables taken to learning for building the predictive model hence, the variable should be relevant to the target concept of the leering process and these variables should not be redundant. Identifying the relevancy and redundancy of variables is called as variable selection. This is a preprocessing stage of knowledge discovery in prediction. Most of the variable selection processes are performed by some statistical or mathematical measures. This study presents a novel way of selecting the variables form the training dataset using unsupervised learners for enhancing the performance of the supervised learners in terms of increasing accuracy and reduce time taken to build the predictive model. The performance of this algorithm is evaluated with fourteen dataset with predictors namely Naive Bayes (NB), Instance Based (IB1) and tree based J48.

Key words: Variable selection, reducing dimensionality, supervised learning, unsupervised learning, ranking, EM clustering, predictive model

INTRODUCTION

The unsupervised learns are known as the classifiers or predictors predict the unlabeled data or unseen data by the predictive model built from learning the historical data (Zhu *et al*., 2010). This predictors research for numerous applications. Now-a-days, improving the performance of this predictor is a great challenge to the researchers. The performance of the predictors highly depends on the significant variables present in the training dataset (Padhy *et al*., 2012). Therefore, the irrelevant and redundant variables must be removed from the training dataset in order to improve performance of the predictor in terms of improving predictive accuracy and reduce the time taken to build the predictive model (Yan *et al*., 2012). This process is called as variable, attribute or variable selection (Sotoca ana Pla, 2010).

The variable selection process is categorized into Filter, Wrapper, Embedded and Hybrid Methods. The Filter Method chooses the significant variables and eliminates the irrelevant and redundant variables form the training dataset based on any mathematical measure without influence of predictors hence, it can be employed for any kind of learning algorithm (Unler *et al*., 2011). The Embedded Method utilizes the training process of the Prediction algorithm to identify the significant variables of the training dataset hence, it depend on the predictors (Haury *et al*., 2011). The wrapper technique utilizes the accuracy measure of the predictors to identify the significant variables present in the dataset (Tsai, 2009). The Hybrid Method combines the concepts of wrapper and filter techniques (Awada *et al*., 2012).

Many of the filter based variable selection approaches concentrate only on selecting relevant variables to the target concepts and fail to remove the redundant variables form the training dataset and fail to identify the redundancy among the variables. This reduces the accuracy of the predictors and increases the time taken to build the predictive model (Gay *et al*., 2010). This study presents a filter based unsupervised Variable Selection algorithm to identify the most significant variables and to remove the irrelevant and redundant variables from the training dataset to improve the performance of the predictors. This algorithm initially clusters the variables by the Clustering algorithms namely K-Means (KM), Farthest First (FF) and Expectation

Corresponding Author: D. Asir Antony Gnana Singh, Department of CSE, Bharathidasan Institute of Technology, Anna University, Tiruchirappalli, India

Maximization clustering (EM) then the clustered variables are selected by information gain measure with a threshold function.

Literature review: In this study, the Variable Selection algorithm, clustering techniques and Prediction algorithms are discoursed as the related works for the construction of the proposed unsupervised Variable Selection algorithm.

Variable Selection algorithm: The Variable Selection algorithm choosing the variables form the training dataset in three fashions: variable subset selection, variable ranking and unsupervised variable selection. In variable subset selection, full set of variables 'F' is divided in to maximum number of possible combinations of variable subset 'S' and these subsets are evaluated based on the evaluation criteria by any one of the mathematical measures in order to select the candidate variable subset. In variable ranking, the significance of individual variable 'f' is measured by mathematical measures and the variables are ranked and selected based on the threshold value. In unsupervised variable selection, the clustering technique is adapted to group the variables in to various clusters and ranking method is applied on each cluster in order to identify the significant variables form the each cluster.

Clustering: Clustering is a technique to group similar objects based on certain criteria such as distance, density, etc. The objects within a group are highly similar than the inter group objects. In Variable Selection algorithms, this fact is applied to group the variables to identify their significance in a training dataset. Many Clustering algorithms are proposed and practiced in various applications based on their purpose and each one has merits and demerits (Kriegel *et al*., 2009).

K-Means clustering (KM): This is an effortless clustering technique compared to other clustering techniques uses the distance measures as the similarity measure for grouping the data objects. Randomly the mean of the cluster is fixed among the data objects and the distance from the mean to the entire data object is calculated by Euclidean distance as expressed in Eq. 1 where the p and q are the two different data objects. The data objects with in a distance from the mean value are grouped as a cluster. The dataset D consist of n data objects and the center point of the cluster is moved and updated in an iterative manner until achieve the perfect cluster of data objects (Jain, 2010):

$$d(p-q)=\sqrt{\sum_{i=1}^{k}(q_i - p_i)} \quad (1)$$

Farthest First clustering (FF): This is an alternative method of K-means clustering. This method adapts the Farthest First Traversal algorithm for fixing up the center point in a specified manner to speed-up the formation of the clusters (Zhao *et al*., 2012).

EM clustering (EM): The data objects are grouped in to various clusters based on their similarity by applying the Expectation Maximization (EM) with maximizing the Maximum Likelihood function (ML) on the data objects as a Gaussian Mixture Model (Mandel *et al*., 2010). The expected similarity of the data objects is computed by Eq. 2 based on this similarity measure the clusters are formed:

$$P(G_Q|F_x,\wp_R)=\frac{P(F_x|G_Q,\wp_R)}{P(F_x|\wp_R)} \quad (2)$$

Predictors: The Prediction algorithm learns the training dataset 'D' and builds the predictive model 'M' for predicting the unlabeled or unseen data. This model is constructed based on various methods such as tree based, rule based and probabilistic based, etc. This study, uses the probabilistic based Naive Bayes (NB), tree based C4.5/J48 (J48) and Instance Based (IB1) predictors.

Naive Bayes (NB): This classifier works with the probability Bayes theory as shown in Eq. 1. The training dataset D contains N number of variables $F = \{f_1, f_2, ..., f_x\}$ and the Classes $C = \{c_1, c_1, ..., c_z\}$. This classifier predicts class C_i for given unlabeled record R by the condition $P(C_k|R) > P(C_w|R)$ where, $k \neq w$ and $n \geq w \geq 1$ known as maximum posterior probability:

$$P(C_k|R)=\frac{P(R|C_k)P(C_k)}{P(R)} \quad (3)$$

where, k = 1, 2, ..., n (Singh *et al*., 2012a, b).

C4.5 (J48): This classifier uses decision tree to build the predictive model to predict the unlabeled record R. This decision tree is constructed by any one of the mathematical measures. Normally, the information gain statistical measure is used in the J48 java implementation of C4.5 as shown in the study. This measure is used to identify most significant variables to split up the training dataset DS. The most significant variable F is chosen as a root node to construct the decision tree. The unlabeled record R is predicted by the decision tree (Polat and Gunes, 2009; Chawla *et al*., 2002).

IB1: This is the instance based classifier that uses the Nearest Neighbor algorithm. This similarity measures is used to predict the unlabeled record R. The Euclidean distance measure is used to calculate the distance between the two records or instances as shown in the Eq. 4 (Ganiz *et al*., 2011):

$$d(R_i, R_j) = \sqrt{\sum_{k=1}^{m} f(R_{ik}, R_{jk})} \quad (4)$$

Where:

$d(R_i, R_j)$ = The distance between the two records R_i and R_j

k = 1, 2, ..., m = The position of the variable F

Proposed algorithm: The proposed Variable Selection algorithm is constructed based on the clustering technique the dataset is fed into Clustering algorithm and the variables of the dataset are clustered based on their similarity. The information gain measure as expressed in Eq. 4 is applied on clustered variables to identify and rank the variables based on their significance. Then, the threshold T_V is applied to cutoff the high significant variables from the ranked variables.

Algorithm: The algorithm of the proposed system is directed as follows:

Input: Training Dataset D
//D consist of set of variables V = {v_1, ..., v_n}
// D consist of set of objects O = {o_1, ..., o_m}
Output: Set of selected variables V_S.

1. Begin;
2. Read D;
3. V_C = Cluster(D) //Clustering the variable
// by KM, EM and FF
4. For [V_{C1}, ..., V_{cn}]
{V_R= Info-gain (V_C)}
//applying information gain on
//clustered variable V_C where V_R ranked
//variable
5. T_V = Compute (V_R)
// computing the threshold value T_V
6. For [V_{C1},...,V_{Cn}]
V_S = Upend (Cutoff (T_v, V_C))
// Vs are the selected variables
7. End

The Information Gain (Info-Gain) measure is formulated in Eq. 5:

$$\text{Info-Gain} = H(Y)-H(Y|X) = H(X)-H(X|Y) \quad (5)$$

The information gained about Y after observing X is equal to the information gained about X after observing Y. The threshold value is calculated by the function for selecting the first half portion of highly significant variables of entire ranked variables.

Table 1: List of datasets

Datasets	Variables in number	Instances in number	Classes in number
Breast cancer	9	286	2
Contact lenses	24	5	5
Credit-g	20	1000	2
Diabetes	8	768	2
Glass	9	214	7
Ionosphere	34	351	2
Iris 2D	2	150	3
Iris	4	150	3
Labor	16	57	2
Segment challenge	19	1500	7
Soybean	683	36	14
Vote	435	17	3
Weather nominal	4	14	2
Weather numeric	14	5	2

MATERIALS AND METHODS

This experiment is conducted with the 14 well known datasets with the number of variables range from 4-683, number of instances range from 5-1500 and the number of class labels range from 2-14 shown in Table 1 are collected from the UCI repository (Bache and Lichman, 2013) and weka datasets (Hall *et al*., 2009). The weka data mining software is installed in the computer system with the specification of Processor: Intel® Core™ 2 CPU T5300 at the rate of 1.73 GHz 1.73 GHz, Memory (RAM): 2550 MB and Operating System: 32 bit Windows Vista home premium to analyze and compare the performance of the proposed Variable Selection algorithm in terms of variable reduction, accuracy produced by predictor and time taken to build the predictive model with that of the existing variable selection methods namely K-means, Farthest First and Maximum Expectation Clustering with the predictors NB, J48 and IBI.

RESULTS AND DISCUSSION

Initially the variable reduction experiment is conducted by K-means, Farthest First and EM Clustering on 14 datasets as seen in Table 1 and the result are shown in Fig. 1-3. Experiment is conducted with the specification of 3 clusters mode. The selected variable subsets are fed in the three predictors namely NB, J48 and IB1 then the predictive accuracy in percentage produced by the predictor with the test mode of 10 fold cross-validation and the time taken to build the predictive model in second are noted.

The performance of the variable selection algorithm in terms of variable reduction, accuracy produces by predictors and time taken to build the predictive model are evaluated. In the analysis on variable reduction as the result shown in Fig. 1. The KM reduces the variable much compared to the EM and FF. In the analysis on accuracy

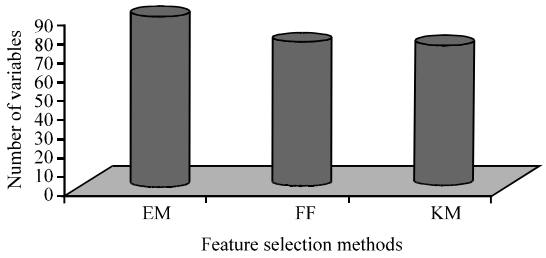

Fig. 1: Comparison on number of variable reduction by the respective Variable Selection algorithm

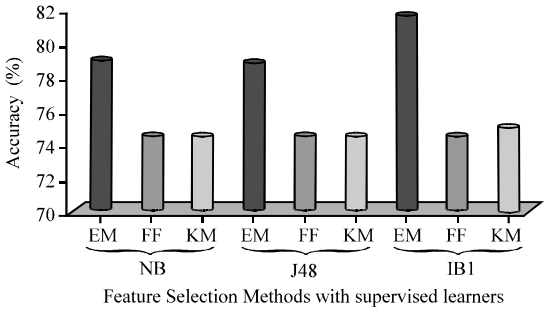

Fig. 2: Comparison on accuracy produced by the predictors with respect to the Variable Selection algorithm

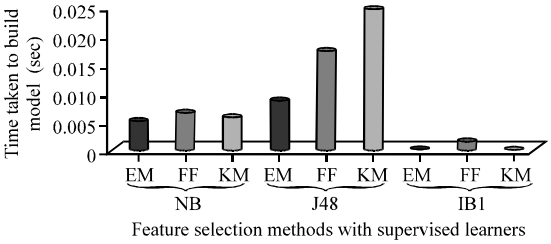

Fig. 3: Comparison on time taken to build the predicitive model by respective Variable Selection algorithms

of predictors as shown in Fig. 2, it is observed that EM produces high accuracy for all the predictors IB1, J48 and NB compare to FF and KM. The KM produces high accuracy with IB1. In analysis on time taken to build the predictive model, as the result shown in Fig. 3. EM spends very less time to build the predictive model for the three predictors IB1, J48 and NB compare to other methods KM and FF. The KM takes less time to build model with IB1 predictor compared to the FF Method.

CONCLUSION

This study presented a unsupervised variable selection for supervised learners and analyses the performance of three unsupervised Variable Selection algorithm KM, FF and EM in terms of variable reduction, accuracy produced by the predictors and time taken to build the Predictive Model. The EM gives overall high accuracy compared to all other Variable Selection algorithms for IB1, J48 and NB predictors. EM takes very less time to build the predictive model for IB1, J48 and NB predictor, in dimensionality reduction the KM perform well compared to the EM and NB. In future, this research can be extended with high dimensional datasets.

REFERENCES

Awada, Wael, T.M. Khoshgoftaar, D. Dittman, R. Wald and A. Napolitano, 2012. A review of the stability of variable selection techniques for bioinformatics data. Proceedings of IEEE 13th International Conference on Information Reuse and Integration, August 8-10, 2012, Las Vegas, NV, USA., pp: 356-363.

Bache, K. and M. Lichman, 2013. UCI machine learning repository. University of California, School of Information and Computer Science, Irvine, CA. http://archive.ics.uci.edu/ml/citation_policy.html.

Chawla, N.V., K.W. Bowyer, L.O. Hall and W.P. Kegelmeyer, 2002. SMOTE: Synthetic minority Over-sampling technique. J. Artificial Intell. Res., 16: 321-357.

Ganiz, M.C., C. George and W.M. Pottenger, 2011. Higher order naive bayes: A novel Non-IID approach to text classification. IEEE Trans. Knowledge Data Eng., 23: 1022-1034.

Gay, G., T. Menzies, M. Davies and K. Gundy-Burlet, 2010. Automatically finding the control variables for complex system behavior. Automated Software Eng., 17: 439-468.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I.H. Witten, 2009. The WEKA data mining software: An update. SIGKDD Explorations Newslett., 11: 10-18.

Haury, A.C. and P. Gestraud and J.P. Vert, 2011. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. PloS one, Vol. 6. 10.1371/journal.pone.0028210.

Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern Recognit. Lett., 31: 651-666.

Kriegel, H.P., P. Kroger and A. Zimek, 2009. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering and correlation clustering. ACM Trans. Knowledge Discovery Data, Vol. 3. 10.1145/1497577.1497578.

Mandel, M.I., R.J. Weiss and D. Ellis, 2010. Model-based expectation-maximization source separation and localization. IEEE Trans. Audio, Speech Language Process., 18: 382-394.

Padhy, N., P. Mishra and R. Panigrahi, 2012. The survey of data mining applications and feature scope. Int. J. Comput. Sci., Eng. Inform. Technol., 2: 43-58.

Polat, K. and S. Gunes, 2009. A novel hybrid intelligent method based on C4. 5 decision tree classifier and one-against-all approach for multi-class classification problems. Expert Syst. Appl., 36: 1587-1592.

Singh, A.G., D. Asir, A. Balamurugan, S. Appavu and E.J. Leavline, 2012a. An empirical study on dimensionality reduction and improvement of classification accuracy using variable subset selection and ranking. Proceedings of the International Conference on Emerging Trends in Science, Engineering and Technology, December 13-14, 2012, Tiruchirappalli, Tamilnadu, India, pp: 102-108.

Singh, D.A.A.G., S.A.A. Balamurugan and E.J. Leavline, 2012b. Towards higher accuracy in supervised learning and dimensionality reduction by attribute subset selection-A pragmatic analysis. Proceedings of the IEEE International Conference on Advanced Communication Control and Computing Technologies, August 23-25, 2012, Ramanathapuram, pp: 125-130.

Sotoca, J.M. and F. Pla, 2010. Supervised feature selection by clustering using conditional mutual information-based distances. Pattern Recognit., 43: 2068-2081.

Tsai, C.F., 2009. Variable selection in bankruptcy prediction. Knowledge-Based Syst., 22: 120-127.

Unler, A., A. Murat and R.B. Chinnam, 2011. *Mr^2PSO*: A maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification. Inform. Sci., 181: 4625-4641.

Yan, T., D. Chu, D. Ganesan, A. Kansal and J. Liu, 2012. Fast app launching for mobile devices using predictive user context. Proceedings of the 10th International Conference on Mobile Systems, Applications and Services, June 25-29, 2012, UK., pp: 113-126.

Zhao, W., Q. He, H. Ma and Z. Shi, 2012. Effective semi-supervised document clustering via active learning with instance-level constraints. Knowl. Inform. Syst., 30: 569-587.

Zhu, X., P. Zhang, X. Lin and Y. Shi, 2010. Active learning from stream data using optimal weight classifier ensemble. IEEE Trans. Syst. Man Cybernetics Part B: Cybernetics, 40: 1607-1621.