Feature Selection using Backward Elimination Iterative Distance Correlation method for adaptive threshold determination

Vadivu.T [1], B.Sumathi [2]
[1]Research Scholar, [2]Associate Professor, Department of Computer Science
CMS College of Science and Commerce

Abstract:
Feature selection is essential for large datasets to eliminate irrelevant and redundant data. Various optimization techniques is used in Software defined networking. This is commonly used method in machine learning. Backward Elimination Iterative Distance Correlation (BE-IDC) method is proposed to solve the fixed threshold values for clustering based feature selection. It is one of the unsupervised cluster based feature selection method for adaptive threshold determination. BE-IDC automatically and dynamically generates threshold values which improves the accuracy of the system.

Keywords: SDN, Feature Selection, Backward Elimination Iterative Distance Correlation.

1. Introduction:

Software Defined Networking (SDN) is designed to manage and control network traffic. This architecture provides support for virtual mobility physical networks. The network control decouples [2] and forwards function by enabling the control to be directly programmable. The amount of traffic in networks is increased with the technologies. It becomes a challenges for detection of anomalies in huge network. SDN simplifies the network design because the commands are listed by controllers rather than protocols.

This framework automatically manages and control huge number of network devices, traffic paths and packet handling using applications interface. Zhang.Y [6] use a prediction based adaptive linear method to detect the anomalies. Unsupervised cluster based feature selection method is used to determine the threshold values in SDN. Feature selection (FS) is a common approach [3] used with machine learning. The subset features are randomly selected for eliminating data. Selection of a subset from existing features is done without a transformation. Feature subset selection method is needed for evaluating a large number of features. Search strategy is used

to find the subsets and to evaluate with the objective function. Filter and Wrapper are two methods of Objective function. Filter evaluates their features with information content. Classifier is used to predict the accuracy for the subset with Wrapper method.

Set of objects in a same group is termed as Cluster. In the existing system [7], two refined algorithms are used such as Density Peak based clustering algorithm (DP) with sample adaption and unsupervised cluster based feature selection mechanism. These algorithms are used to extract the cluster centers and outliers location automatically. The relevant features were clustered into group based on their maximum redundancy to remove irrelevant features. Two fixed threshold values are used for selecting and removing redundant features. Fixed values may not provide clear extraction of relevant features. DP is appropriate for static dataset only. Intrusions are not defended timely. Classification accuracy has not increased and it takes more time while processing a huge amount of data. Backward Elimination Iterative Distance Correlation method (BE- IDC) is proposed to solve fixed threshold issue for cluster feature selection algorithm. BE-IDC automatically sets threshold values. Classification accuracy is sufficient with Dynamic generation of subset features without any prior knowledge. Feature selection is used with machine learning to train up faster. It reduces the complexity of the model and improves accuracy. This method removes the

irrelevant, redundant or noisy data from the subset feature.

2. Methodology:

Feature selection (FS) is defined as the process of identifying and removing the irrelevant and redundant information. Feature is also called as variable or attribute of the data. This can be classified as relevant, irrelevant and redundant. Feature selection is a commonly used process in machine learning algorithm. This is a process which chooses an optimal subset of features according to a certain criteria. This method searches the finest subset feature among all the features. The best subset consists the near to the ground proportion which leads to accuracy. Rest features are discarded. The main goal of the feature selection is to identify a minimal feature subset. The need for the feature selection is to improve the performance with speed, predictive power and simplicity of the subset. It improves the storage requirements. When using this method the data quality is amplified. The subset decrease is useful to save the resources for the consumption of next iteration. When subset feature SF={ SFi|i=1…N}, then find a subset $S_{x,}$ with x<y, which maximizes an objective function OF(S). So S_x ={ SF_{i1}, SF_{i2}, ….SF_{ix}}=arg max OF=={ SFi|i=1…N}.A search criteria is required to search all the possible combinations of features. The Evaluation measures include probability of error, divergence, dependence, interclass of distance, consistency [3]. The feature selection algorithm is shown below:

General Algorithm: Feature Subset Selection

Input:

F - data with features S,|S|=n

M- evaluation measure to be maximized

GS – generation successor operator

Output:

Key - subset feature

N:= Begin_Position (S)

Key := { best of N according to M};

repeat

N := Find_Criteria (N,GS(M), S);

S' = { best of N according to M};

If M(S') ≥ M (Key) or M(S') =M(Key)

and |S| ‹|Key| then Key := S'

until end (M,N)

Classification of feature selection includes Filters and Wrapper. In filter method, the evaluation is done by using their data content. The wrapper method uses a classifier to identify subsets for predictive accuracy. Forward elimination and backward elimination are two classifications of wrapper method. Backward feature elimination technique is used to remove the redundant data from subset. The design of wrapper feature selection method is shown in fig 1.

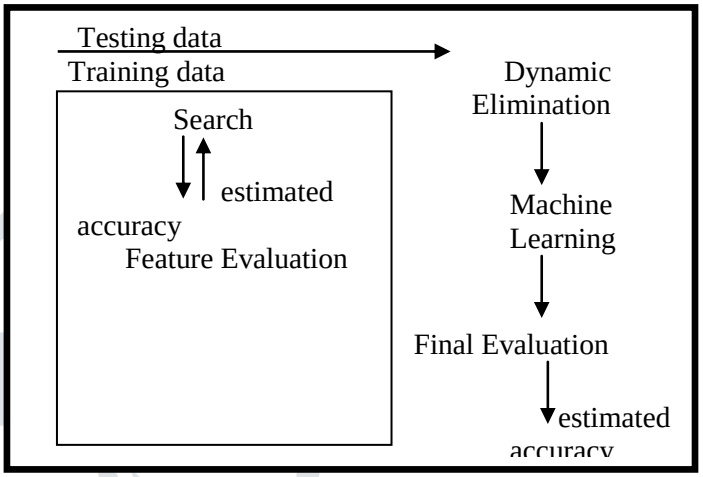

Fig 1. Wrapper feature selection

2.1 Backward Elimination Iterative Distance Correlation (BE-IDC):

This is a wrapper method which uses a classifier to assess subset feature. BE-IDC is an unsupervised cluster based feature selection method. This technique automatically determines the threshold values randomly. BE works in the reverse direction. This is also called as Sequential Backward Elimination. This process starts with set of all variables, iteratively removes the features which results in lowest value of the objective function. Elimination of redundant features results in decrease of objective function. The evaluation measure depends on probability, consistency and accuracy. BE spends most of the time by visiting its subsets. The subset consists of unique features. There is an inability to reevaluate the usefulness of a feature when it has been discarded. The backward elimination iteration selection is in shown fig 2.

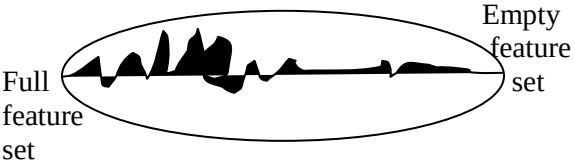

fig 2. Backward Elimination

a. Sequential Backward Elimination Feature Algorithm

BE algorithm initially starts with full feature subset generation. This sequentially eliminates the feature subset, reduces the value of the objective function. BE algorithm is shown below:

Algorithm: Sequential Backward Elimination Feature Set Generation

```
function SBG (F- full set, M- measure, S-holds the removed features)
        initialize :S= { }
repeat
        x=GetNext(F)
        F=F-{x}
        S=SM{f}
until S does not satisfy M
return FM{x}
end function
```

The function begins with full feature subset F generated as SBG. The evaluation measures M sequentially eliminates the subset features x. The eliminated features S actually increase the value of the objective function. All the features are iteratively repeated until it removes all the worst subset features. The random generation of subset feature is shown below:

Algorithm: Random feature set generation RG

```
function RG (F- full set, M-measure)
        begin : S= S best = {  }
        begin : C best  = #(F)
repeat
        S = RANDGEN (F)
        C = #(S)
        If C ≤ C best and S satisfies M then
                S best = S
                C best = C
        endif
        until the condition is satisfied
        return subset
end function
```

Random feature generation starts with full subset S and randomly generates the features RG. Initialize all the attributes for the function F. Then repeat the process of random generation RANDGEN for a function, until the best subset is found. Once the condition is satisfied, end the function.

b. Backward Elimination Scheme

Identifying a threshold value separates the normal features from noise feature. This Backward elimination iterative distance correlation method [9] automatically identifies an optimal threshold value for prediction accuracy. D denotes the threshold value to be determined. S_C = {S_x, x=1,…, d} be a subset for the function. S_N = { S_x, x=d+1,…, J} be a set of noise features. The goal of the elimination is to remove the redundant feature S_N and to collect subset S_C. Initialize from the largest subset X_J, which consists of all the recorded features datasets. Then the noise features are discarded by backward elimination with much iteration. The prediction based error is calculated for each iteration in the dataset.

Step 1: All the features are ranked and recorded set is obtained X_J, where J is expected to be higher.

Step 2: Begin from the largest dataset X_J and calculate the prediction error for the related subset.

Step 3: Least correlation values is removed with a particular fall rate of feature based on the ranks through step 1. Calculate prediction error for the current subset.

Step 4: Repeat step 3 is until the least feature (min) is identified.

Step 5: We can plot and locate the model size with the number of features. So, that the selected feature set S_C and adaptive threshold d can be determined. The noisy features are thrown away during the iterations of step 3 & 4.

The prediction accuracy is accurate using this BE method. For example, run 100 bootstrap samples of feature selection. Each bootstrap sample is divided into training data and test data. The accuracy of the system is computed using a classifier technique. The original accuracy is compared with normal feature. Remove each attribute by using the classifier one at a time and find the accuracy. The absolute deviation and the test data are compared. If new value is higher than the normal then move to new subset, else move to the original value. It is repeated until no improvements in the values are identified or full subset attributes have been done in. The successive feature selection of backward elimination scheme is shown in fig 3.

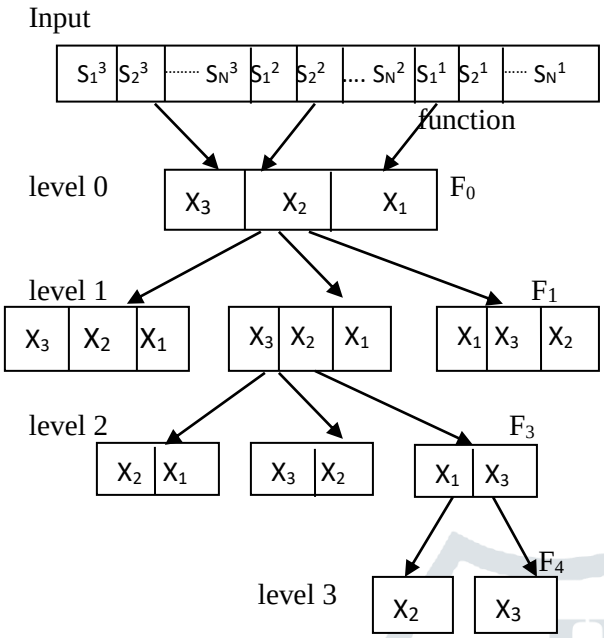

Fig 3. Successive feature selection

3. Conclusion

BE-IDC method is proposed to fix the threshold values for selecting and removing redundant features. This has a better performance with optimal feature subset, BE spends most of its time by visiting the large subsets. This method enables higher prediction accuracy. Adaptive threshold is efficient for separation of noise features from the subset. Future extension of this technique is to propose an optimized BE-IDC for feature selection and clustering process without any prior knowledge.

References:

1. Zhang, Y. (2013), "An adaptive flow counting method for anomaly detection in SDN", Proceedings of the ninth ACM conference on Emerging networking experiments and technologies, (pp. 25-30).
2. Vadivu.T, B.Sumathi, "A survey on Software Defined Networking with traffic Anomaly Detection Techniques", International Journal of Research and Analytical reviews, 2348-1269.
3. L.Ladha, T.Deepa, "Feature Selection methods and Algorithms", International journal of Science and Engineering, Vol 3, 2011.
4. Zhong W, Zhu L, "An iterative approach to distance correlation-based sure independence screening", JStat Comput Simul, 2014; 85(11):2331–45.
5. Guifang Fu, Gang Wang, Xiaotian Dai, "An adaptive threshold determination method of feature screening for genomic selection", (2017),Mathematics and Statistics Faculty Publications, Paper 218.
6. Zhang, Y. (2013), "An adaptive flow counting method for anomaly detection in SDN", In Proceedings of the ninth ACM conference on Emerging networking experiments and technologies, (pp. 25-30).
7. Xia, W., Wen, Y., Foh, C. H., Niyato, D., & Xie, H. (2015), " A survey on software-defined networking. IEEE Communications Surveys & Tutorials", 17(1), 27-51.
8. Shailendra Singh, Sanjay Silakari, "An ensemble approach for feature selection of Cyber Attack Dataset", Department of Information Technology, Bhopal, India., Vol 6, Nov 200.
9. Fu, Guifang; Wang, Gang; and Dai, Xiaotian, "An adaptive threshold determination method of feature screening for genomic selection", 2017, Mathematics and Statistics Faculty Publications, Paper 218.
10.Study on Feature Selection Techniques in Educational Data Mining, M. Ramaswami and R. Bhaskaran,Vol1, Dec, 2009.