Hybrid Method Based Feature Selection using Simulated Annealing and Fuzzy Clustering Techniques

1Aneetha A.S. and 2Bose S.

1Department of Computer Science and Engineering, Anna University, Chennai, 600025, Tamil Nadu, India.
2Associate Professor, Department of Computer Science and Engineering, Anna University, Chennai, 600025, Tamil Nadu, India.

ARTICLE INFO
Article history:
Received 12 October 2014
Received in revised form 26 December 2014
Accepted 1 January 2015
Available online 25 February 2015

Keywords:
Feature subset selection,
Hybrid approach,
Fuzzy k means,
Simulated annealing.

ABSTRACT
Feature selection is an important aspect in the field of knowledge discovery and data mining which deals with large amounts of data with relevant and irrelevant features to their target classes. Selecting a good subset with relevant features and eliminating irrelevant ones for classification and pattern recognition problems is a challenging task. In this work a hybrid feature subset selection approach has been proposed for the classification problems based on simulated annealing and fuzzy k-means algorithms. Since the hybrid approach combines both the filter and wrapper approaches, records are randomly selected and feature threshold is used to filter the features which are then given to the Fuzzy k-means (FKM) classifier for evaluation. Simulated annealing (SA) approach is used to repeatedly search the features and evaluate the classifier until optimum accuracy is achieved. The approach is evaluated using KDD cup '99 dataset which has five different classes and that can be used to select the relevant features according to their target class. The results that are reported in this work have been achieved with optimum accuracy by employing comparatively lesser number of features.

© 2015 AENSI Publisher All rights reserved.


INTRODUCTION
As computer and database technologies advance rapidly, volume of data accumulates are unmatchable by human’s capacity of data processing. Feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility [Song et al., 2013]. The choosing a subset of relevant features with respect to the target concepts becomes necessary in all the classification and pattern recognition problems.

Many feature subset selection methods are proposed for machine learning approaches [Liu and Yu, 2005] which are broadly classified into wrapper, filter and hybrid approaches. While the wrapper approach evaluates the accuracy of the predefined classifier with the selected subset of features whereas the filter approach is independent of learning algorithm but is relevant only to data set. Both the approaches involve complex computation as search is performed on the feature space. The hybrid approach is the combination of filter method which reduces the search space and uses the wrapper method for subsequent evaluation. Therefore, optimization here becomes essential. The hybrid model attempts to take advantage of the two models by exploiting their different evaluation criteria in different search stages.

Fuzzy Clustering analysis has been widely used in many areas such as network security, artificial intelligence, pattern recognition, data mining, and image processing [Song et al., 2013][Jain and Rui, 2009][Wang et al., 2010]. It is a way of grouping supervised or unsupervised data into one or more clusters based on the similarity measures. Setting the parameter and selecting relevant features for fuzzy clustering becomes the interesting problem in the area of network security, machine learning etc. In this work we propose the feature selection for the fuzzy k means algorithm based on hybrid feature selection approach. The greedy search technique, simulated annealing, is deployed for identifying the feature subset and fuzzy k means technique is used to evaluate the feature subset.

1. Literature review:
Feature subset selection consists of identifying the relevant features and discarding the irrelevant ones, with the goal of obtaining a subset of features with a minimum degradation of performance [Song et al., 2013][Liu
Many statistical approaches such as factor analysis (FA), principle component analysis (PCA) etc., have been deployed to investigate the performances measures, mostly based on distance and information measures in the field of feature selection [Chou et al., 2008][Lin et al., 2008].

Chou et al., [2008] has proposed simulated annealing based parameter determination and feature selection for the support vector machine (SVM) classifier and also in the same year they have proposed to set the learning rate, number of hidden neurons and momentum along with feature subset selection for the back – propagation networks [Lin et al., 2008]. In both their works they have used simulated annealing to obtain the optimum parameter values and feature subset to achieve the high classification accuracy rate. Lin et al., [2012] has proposed an intelligent algorithm for network anomaly detection based on simulated annealing, SVM and decision tree. They obtain the relevant features using SA + SVM and based on those features they obtain the decision rules for detecting anomalies in the network traffics. The combined approach of simulated annealing and fuzzy clustering are utilized by Filippone et al., [2006] for gene selection in order to reduce the dimension in DNA micro array data.

Many researchers deploy fuzzy clustering in their area of classification domain, where they need to specify number of clusters and its fuzzifier value which is a great challenge to them. Kuo – Lung Wu [2012] has analyzed the selection of parameters for fuzzy c- means using cluster – core base methods, they showed that larger fuzzifier value will make FCM more robust to noise and outliers. Bases on the above literature we have proposed hybrid feature subset selection method by employing simulated annealing techniques to identify relevant and efficient features for the fuzzy k means classifier.

2. Methodology:

In this section hybrid feature selection approach is discussed, which deploys global optimization technique simulated annealing (SA) to perform the selection of most relevant and efficient features for fuzzy k-means (FKM) classifier. A typical feature selection process consists of four basic steps, namely, subset generation, subset evaluation, stopping criterion, and result validation. The Figure 1 shows the process of subset generation and evaluation is repeated until a given stopping criterion is satisfied [Liu and Yu, 2005].

![Fig. 1: Hybrid Feature Subset Selection Approach.](Image)

Simulated annealing approach is an iterative which repeats until optimum solution is reached [Chou et al., 2008][Bolon-candeo et al., 2011]. First initial solution is obtained based on objective function, next new solution is generated by evaluating the objective function with randomly selected new set of features. If the new solution is better than the initial solution then assign new solution as initial solution otherwise Metropolis principle is used to decide the acceptance or rejection of the new solution. The process is repeated until it reaches the stopping criteria, defined initially.

Data Source:

The proposed algorithm is evaluated using the 10% corrected KDD Cup’99 dataset [KDD Cup, 1999] which has different class and each class has their own pattern of traffic profiles. The dataset contains normal and 21 different types of attack classes which are broadly classified into four categories namely Denial of Services (DoS), Probing (Probe), User to Remote (U2R) and Root to Local(R2L). Each profile is represented using 41 features along with one for class representation. As our objective is to select the relevant features, the details about the dataset is important which are available in literatures referred in [Lin et al., 2012][Luo and Xia 2014]. It has 32 continuous and 9 discrete features in that 3 have categorical values and 6 have nominal values. As from earlier literatures we known that dataset has redundant patterns which need to be removed to design an efficient system. The Table 1 below gives the brief description of number of traffic patterns before and after removing redundant records in all the five classes.

<table>
<thead>
<tr>
<th>Classes</th>
<th>Normal</th>
<th>DoS</th>
<th>Probe</th>
<th>R2L</th>
<th>U2R</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of records originally</td>
<td>97,279</td>
<td>3,91,460</td>
<td>3460</td>
<td>442</td>
<td>37</td>
</tr>
<tr>
<td>No. of unique records</td>
<td>87,832</td>
<td>54,573</td>
<td>1627</td>
<td>425</td>
<td>37</td>
</tr>
</tbody>
</table>

Pre-processing:

Since our algorithm performs feature selection, it needs to give uniform importance to each feature irrespective of their feature values. This can be carried out by performing scaling or normalization process, where every feature values are transformed to new values within the predefined range. The efficient Min – Max
normalization technique [Han et al., 2006] is deployed in here to perform the value transformation to the range [0,1]. The Min–Max formula is given as

\[ F_{k,i,new} = \frac{(F_{k,i,old} - F_{i,min})}{(F_{i,max} - F_{i,min})} \]  

where

- \( F_{k,i,new} \): New value of the ith feature of record k.
- \( F_{k,i,old} \): Old value of the ith feature of the record k.
- \( F_{i,min} \): Minimum value of the ith feature.
- \( F_{i,max} \): Maximum value of the ith feature.

**Feature Selection Algorithm:**

The given SA + FKM clustering algorithm searches for optimal parameter value to minimize the objective function of the clustering method [Pal and Bezdek, 1995]. The SA + FKM approach starts with the following five variables: initial temperature (\( T_0 \)), or final temperature (\( T_F \)), number of iterations (\( I_{iter} \)), attenuation constant or annealing schedule (\( \beta \)) and the threshold for stopping the clustering process(\( \theta \)). The initial annealing temperature should be a larger value based on the iterations needed for cooling process and set as \( T \). While solving the minimization problem, the next solution gets accepted as a feasible solution only when the value produced by the objective function is minimum. The searching process continues by keeping the newly accepted point as the current solution and seeks for the next feasible solution. If the next solution gives greater value for the objective function than the current solution Metropolis principle of acceptance rule is applied to find the probability of accepting the greater value as the current solution [Kirkpatrick, 1984]. The proposed approach pseudocode is given in figure 2 and 3.

**RESULTS AND DISCUSSIONS**

The SA + FKM approach is evaluated using 10 fold cross validation technique for evaluating the objective function of the FKM with. The experiments are conducted for ten times, the average performance of cost function is considered as the final results. After completing the pre-processing, algorithm given in the previous section needs to initialize parameters values for conducting the experiments. The values assigned for those parameters are tabulated below in Table 2. As the next step, initial feature subset has to be selected using feature threshold. In this work mean value of the features are considered for threshold setting. As trial and error approach we have assigned different values as feature threshold from 0.05 to 0.5 and the number of features selected for each threshold is given in the Figure 4. The Figure 4 shows the number of features selected for each threshold from that it can be clearly seen that with threshold 0.5 all the datasets have more or less same number of features.

So, 0.5 is considered as filtering criteria to perform the experiments further. After selecting a feature subset, the same is evaluated for optimum cost function as explained in the previous section. The cost function value obtained before and after features selection is given in the Table 3 for all the five different classes.
Fig. 3: Pseudocode of FKM.

Table 2: Parameter values.

<table>
<thead>
<tr>
<th>Meaning</th>
<th>Symbol</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Temperature</td>
<td>T0</td>
<td>1000</td>
</tr>
<tr>
<td>Final temperature</td>
<td>TF</td>
<td>0.00</td>
</tr>
<tr>
<td>Number of iterations</td>
<td>Iter</td>
<td>10</td>
</tr>
<tr>
<td>Attenuation Constant or Annealing Schedule</td>
<td>B</td>
<td>0.8</td>
</tr>
<tr>
<td>Stopping criteria</td>
<td>Θ</td>
<td>0.01</td>
</tr>
<tr>
<td>Feature Threshold</td>
<td>Fm</td>
<td>0.5</td>
</tr>
<tr>
<td>K-fold validation</td>
<td>K</td>
<td>10</td>
</tr>
<tr>
<td>Number of clusters</td>
<td>C</td>
<td>Varied for each class</td>
</tr>
<tr>
<td>Fuzzifier parameter</td>
<td>M</td>
<td>2</td>
</tr>
</tbody>
</table>

Fig. 4: No. of Features Selected Vs Feature Threshold.

Table 3: Performance of proposed algorithm.

<table>
<thead>
<tr>
<th>Classes</th>
<th>Normal</th>
<th>DoS</th>
<th>Probe</th>
<th>R2l</th>
<th>U2R</th>
</tr>
</thead>
<tbody>
<tr>
<td>CF before feature selection</td>
<td>276.6152</td>
<td>16.2592</td>
<td>52.2711</td>
<td>14.8069</td>
<td>18.180</td>
</tr>
<tr>
<td>CF after feature selection</td>
<td>38.6936</td>
<td>0.4743</td>
<td>10.127</td>
<td>1.5529</td>
<td>2.4901</td>
</tr>
</tbody>
</table>

We can understand from Table 3 that the performance of the cost function has been optimized with feature selection. Especially with respect to the DoS data set, the proposed approach is able to achieve very low cost function value approaching 0. The time taken for analysis with feature selection is slightly larger because of more number of iterations has been performed with large number of variables for optimization. The method identifies that 19 features are relevant out of 41 for addressing both attack and normal data for the KDD cup 99 bench mark dataset.

Table 4: No. of Features Selected.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of features selected</td>
<td>19</td>
<td>25</td>
<td>26</td>
</tr>
</tbody>
</table>

The number of features selected by our proposed hybrid approach and different approaches based on simulated annealing is given in Table 4. Therefore, with respect to cost function, our hybrid feature selection method based on SA + FKM clustering approach is found to be better and identifies relevant features for that particular classifier.

Conclusion:

In this paper, we have presented a new SA + FKM based hybrid feature subset selection approach for identifying both relevant and irrelevant features towards the targeted concept. This approach involves 1) random feature selection based on feature threshold 2) evaluating the feature performance in terms of minimum cost
function using FKM and 3) repeating the above two steps until global optimum features are achieved using SA. Minimum cost function has been achieved with only 19 out of 41 features, by performing feature selection on the KDD cup '99 dataset. With feature selection, the approach reported in this work has been found to delete the insignificant features effectively while maintaining better cost function value. When compared to the existing feature selection approaches, the SA + FKM approach is found to give better results in terms of utilizing lesser number of features to achieve minimum cost function. In future we can cascade this SA + FKM approach along with other classical classification or clustering techniques to achieve high accuracy and light weight systems to perform knowledge discovery in various domains.

REFERENCES


Han, J. and M. Kamber, J. Pei, 2006. Data mining: concepts and techniques, Morgan Kaufmann.


