Classification Using the Compact Rule Generation

NAVNEET and NASIB SINGH GILL

Department of Computer Science and Applications, Maharshi Dayanand University, Rohtak Haryana, India.

(Received: March 11, 2015; Accepted: April 18, 2015)

ABSTRACT

Various attributes within a dataset relate to each other and with the class attribute. The relationship between the different attributes with class attribute may improve the classification accuracy. The paper introduces CCSA algorithm that performs the clustering that is cascaded by classification based on association. The clustering process generates a group of various instances within the dataset. These clustered instances are classified by using the association. This paper uses the Apriori association to generate the rules for classification. The technique is analyzed by using the soil data set and various other online available datasets using WEKA. The simulation result using the WEKA shows that reduced rules with the improved classification accuracy as compared to the existing association with classification algorithms.

Key words: Data mining, PART, WEKA, k-mean clustering, Schwarz criteria, Association.

INTRODUCTION

Data mining is the process to get the hidden information available within the database and it can lead to the knowledge discovery\(^1\,^2\). Various operations included in the data mining are the classification, prediction, association and the clustering. The association\(^3\) is most useful application that is used to get the relationship between the elements of particular set. These properties are based on co-occurrence of the data items instead of inherent properties of data. Main motive of association is to extract the frequent patterns, associations, casual structures and interesting correlations among various set of items of different transactions of databases or data repositories. Association is mainly applied to the risk management, telecommunication market etc. Association is used to define the relation between various attributes of the active database to get the frequent pattern and the correlation between various set of items within the dataset. This analysis is helpful to get the exact idea of the customer behavior that leads to the progress in business. The number of association rules is large for any given dataset, so association mining\(^4\) is used to get the association rule for a given dataset. In the association mining the dataset is firstly decomposed...
into the sub sets and this process carries on to get the sets of various items within the dataset. The item sets that exceeds the threshold value is known as large or frequent item set. Then the association rules are build to derive the relation between various item sets. This paper is further divided into four sections section 1 describe the association and PART algorithm. The section 2 discusses the related work and the section 3 describes the proposed technique. The section 4 describes the result with their analysis.

**Apriori algorithm**

Various association algorithms exist to find the relationship among various attributes of the dataset. Apriori algorithm is found to be the most efficient algorithm among them. The Apriori algorithm completes its process in two steps, first step is to generate the items that has support factor greater than or equal to, the minimum support and the second step is to generate all rules with confidence factor greater than or equal to the minimum confidence. In the first step, the set of possible itemsets is exponentially increasing that makes the finding of frequent dataset more difficult. The downward closure property is used to find the frequent item sets. The Apriori algorithm find subsets which are common to threshold number of itemsets and uses a “bottom up” approach. The frequent subsets are extended one item at a time and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found.

**Classification rule based part algorithm**

Classification is a concept or process of finding a model which finds the class of unknown objects. It basically maps the data items into one of the some predefined classes. Classification model generate a set of rules based on the features of the data in the training dataset. Further these rules can be used for classification of future unknown data items. Classification is the one of the most important data mining technique. Medical diagnosis is an important application of classification for example, diagnosis of new patients based on their symptoms by using the classification rules about diseases from known cases.

PART stands for Projective Adaptive Resonance Theory. The input for PART algorithm is the vigilance and distance parameters.

**Initialization**

Number m of nodes in F1 layer := number of dimensions in the input vector. Number m of nodes in F layer := expected maximum number of clusters that can be formed at each clustering level.

Initialize parameters L, ρ₀, ρ₁, α, θ, and e.
1. Set ρ₀ = ρ₀.
2. Repeat steps 3–7 until the stopping condition is satisfied.
3. Set all F2 nodes as being non-committed.
4. For each input vector in dataset S, do steps 4.1–4.6.
4.1. Compute hᵢ for all F1 nodes vᵢ and committed F2 nodes vⱼ. If all F2 nodes are non-committed, go to step 4.3.
4.2. Compute Tⱼ for all committed F2 nodes Vⱼ.
4.3. Select the winning F2 node Vⱼ. If no F2 node can be selected, put the input data into outlier O & then continue to do step 4.
4.4. If the winner is a committed node, compute rⱼ, otherwise go to step 4.6.
4.5. If rⱼ >= ρ, go to step 4.6, otherwise reset the winner Vⱼ and go back to step 4.3.
4.6. Set the winner Vⱼ as the committed and update the bottom-up and top-down weights for winner node Vⱼ.
5. Repeat step 4 N times until stable clusters are formed (i.e. until the difference of output clusters say Nth and (N-1)th time becomes sufficiently small).
6. For each cluster Cⱼ in F2 layer, compute the associated dimension set Dⱼ. Then, set S=Cⱼ and set ρ = ρ+ ph (or ρ = l D l = ph), go back to step 2.
7. For the outlier O, set S = 0, go back to step 2.

**Related work**

Vamanan. R (2011) finds the suitable model to predict the yield production. The author compares MLR and k-mean clustering technique to determine the average production of the rainfall on the dataset containing the data regarding yearly rainfall in cms by using four parameters i.e. Year,
Rainfall, Area of Sowing and Production. The results are compared with the actual average production then MLR technique gives 96% accuracy while the K-MEAN clustering gives the 98% accuracy. Shweta et al. (2012) have implemented clustering technique using WEKA tool to create clusters of the soil based on their salinity. The author creates the cluster to find the relation between the various types of soils in the selected subset of soil. The subset of soil is selected from the world soil science database. D Ramesh et al. (2013) used k-means approach to estimate the crop yield analysis. They also reviewed various methodologies that are used in agricultural domain and finds Naïve Bayes classifier suitable for the soil classification. Navneet et al. (2014) used Schwarz Criterion (SC) to choose the optimal number of clusters in the k-mean algorithm for a given range of values according to intrinsic properties of the specified data set. Schwarz Criterion is a parametric measure of how well a given model predicts the data. It represents a trade-off between the likelihood of the data for the model and the complexity of the model. The presented algorithm is implemented using the WEKA tool and analyzed on various datasets available at internet. The accuracy of 97.17% is found on the soil dataset.

**CCSA Technique**

The paper already showed that higher accuracy can be achieved by cascading the clustering with the classification. Moreover, in the paper the technique uses the Schwarz criteria to decide the number of clusters. The technique in the paper gives better results than the existing classification technique. This technique minimizes the classification rules by using the relationship between the attributes instead of gain ratio. The Boolean relation is used to classify the instances. It can also improve the classification accuracy. In CCSA technique (Cascading of Clustering based on Schwarz criteria and Association) the elements within the dataset are divided in to the number of clusters. The number of clusters within the dataset is determined by the Schwarz criterion. Then the elements within the clusters are classified. The classification occur process determine the frequency of the each class within the clusters. The low frequency component joined with the other components to make a frequent set. Then

![Fig. 1: Block diagram of CCSA Technique](image)
the relation is derived between various sets of the clusters. This relation is reduced as the number of sets is reduced due to combining various components in a set. This process is applied to each cluster. The process is explained in the following block diagram.

The process briefed in the above block diagram can be explained by the following algorithm:

**CCSA Algorithm**

I. Input large Dataset of soil sample.

II. Initiate K=smallest value(default k=2);

III. Apply K-means to generate number of clusters say C0, C1, C2, .... , Cn.

IV. For i=1:n

V. Calculate the Schwarz criterion for cluster Ci by using

$$SC = -2 \ln \hat{L} + k \ln(n)$$  \hspace{1cm} (1)

Where \( \hat{L} \) = data within the cluster Ci
n= the number elements in Ci
k = the number of parameters to be estimated.

\( \ln \) = The maximized value of the likelihood function of the model M i.e \( \hat{L} = p(x|\hat{\beta}, M) \) where \( \hat{\beta} \) are the parameter values that maximize the likelihood function.

VI. Apply K-mean on Ci Clusters for k=2 say generated Clusters are Ci1 and Ci2

VII. Calculate the SC for Clusters Ci1 and Ci2 by using

$$SC_1 = -2 \ln \hat{L} + 2 * k \ln(n)$$  \hspace{1cm} (2)

Here, the number of parameters get doubled due to two cluster.

VIII. If SC>SC1 then n=n+1 i.e. new model preferred.

IX. Ci=Ci1 and Cn=Ci2

X. i=i-1

XI. End if

XII. End

XIII. For each cluster

XIV. S=elements of Selected Cluster

XV. l=1

XVI. While Li is not empty

XVII. Se(i+1)=elements generated by Li

XVIII. For each element in the cluster

XIX. Count the candidates of the Se(i+1)

XX. S(i+1)=candidates of Se(i+1) with minimum support

XXI. End

XXII. End

The above process can be implemented by using the WEKA. This process is analyzed on the soil dataset explained in our previous paper. This algorithm is also analyzed on the diabetes and ionosphere dataset. The analysis of these

### Table 1: Comparison of Various algorithms using various parameters

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Number of rules</th>
<th>Classification accuracy</th>
<th>TP rate</th>
<th>FP rate</th>
<th>Precision</th>
<th>recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>PART</td>
<td>40</td>
<td>94.69</td>
<td>0.947</td>
<td>0.047</td>
<td>0.944</td>
<td>0.947</td>
<td>0.945</td>
</tr>
<tr>
<td>K Mean + Association</td>
<td>24</td>
<td>96.19</td>
<td>0.952</td>
<td>0.045</td>
<td>0.950</td>
<td>0.952</td>
<td>0.950</td>
</tr>
<tr>
<td>Proposed(CCSA)</td>
<td>13</td>
<td>97.69</td>
<td>0.956</td>
<td>0.042</td>
<td>0.955</td>
<td>0.956</td>
<td>0.955</td>
</tr>
</tbody>
</table>

### Table 2: Comparison of PART, Kmean+Association, proposed(CCSA) algorithms on different datasets

<table>
<thead>
<tr>
<th>Name</th>
<th>Number of instances</th>
<th>Number of attributes</th>
<th>PART</th>
<th>Kmean+ Association</th>
<th>Proposed (CCSA)</th>
<th>Classification accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetes</td>
<td>768</td>
<td>9</td>
<td>13</td>
<td>6</td>
<td>3</td>
<td>75.26</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>351</td>
<td>35</td>
<td>10</td>
<td>7</td>
<td>4</td>
<td>91.73</td>
</tr>
</tbody>
</table>

The above process can be implemented by using the WEKA. This process is analyzed on the soil dataset explained in our previous paper. This algorithm is also analyzed on the diabetes and ionosphere dataset. The analysis of these
and comparison of results with the PART algorithm and our existing and the kmean clustering with association based classification is done in the next section.

RESULTS AND ANALYSIS

The simulation of the CCSA algorithm is done using the WEKA. The comparison is done between the PART algorithm and the Kmean clustering with association

Rules generated by the CCSA techniques are as follow:

\[
\begin{align*}
  &p\text{Clust}_0 < 0.000023 \quad \text{AND} \\
  &p\text{Clust}_1 < 0.000019: \text{MLH (639.0/4.0)} \\
  &p\text{Clust}_2 < 0.004404 \quad \text{AND} \\
  &p\text{Clust}_3 < 0.018951 \quad \text{AND} \\
  &p\text{Clust}_4 < 0.005588: \text{MLM (311.0/30.0)} \\
  &p\text{Clust}_5 < 0.039564 \quad \text{AND} \\
  &p\text{Clust}_6 < 0.000091 \quad \text{AND} \\
  &p\text{Clust}_7 < 0.000052: \text{MLM (506.0/12.0)} \\
  &p\text{Clust}_8 < 0.036364: \text{MLM (233.0/29.0)} \\
  &p\text{Clust}_9 < 0.09236 \quad \text{AND} \\
  &p\text{Clust}_{10} < 0.09236 \quad \text{AND} \\
  &p\text{Clust}_{11} < 0.000177: \text{MLM (560.0/3.0)} \\
  &p\text{Clust}_{12} < 0.206039 \quad \text{AND} \\
  &p\text{Clust}_{13} < 0.206039 \quad \text{AND} \\
  &p\text{Clust}_{14} < 0.206039 \quad \text{AND} \\
  &p\text{Clust}_{15} < 0.206039 \quad \text{AND} \\
  &p\text{Clust}_{16} < 0.206039 \quad \text{AND} \\
  &p\text{Clust}_{17} < 0.206039 \quad \text{AND} \\
  &p\text{Clust}_{18} < 0.206039 \quad \text{AND} \\
  &p\text{Clust}_{19} < 0.206039 \quad \text{AND} \\
  &p\text{Clust}_{20} < 0.000177: \text{MLM (560.0/3.0)} \\
  &p\text{Clust}_{21} < 0.000173 \quad \text{AND} \\
  &p\text{Clust}_{22} < 0.000091: \text{MLM (28.0)}
\end{align*}
\]

Rules generated by the existing techniques are as follow:

\[
\begin{align*}
  &p\text{Clust}_0 < 0.000024 \quad \text{AND} \\
  &p\text{Clust}_1 < 0.000024 \quad \text{AND} \\
  &p\text{Clust}_2 < 0.000024 \quad \text{AND} \\
  &p\text{Clust}_3 < 0.000024 \quad \text{AND} \\
  &p\text{Clust}_4 < 0.000024: \text{MLM (9.0/1.0)} \\
  &p\text{Clust}_5 < 0.000024 \quad \text{AND} \\
  &p\text{Clust}_6 < 0.000024 \quad \text{AND} \\
  &p\text{Clust}_7 < 0.000024 \quad \text{AND} \\
  &p\text{Clust}_8 < 0.000024 \quad \text{AND} \\
  &p\text{Clust}_9 < 0.000024 \quad \text{AND} \\
  &p\text{Clust}_{10} < 0.000024 \quad \text{AND} \\
  &p\text{Clust}_{11} < 0.000024 \quad \text{AND} \\
  &p\text{Clust}_{12} < 0.000024 \quad \text{AND} \\
  &p\text{Clust}_{13} < 0.000024 \quad \text{AND} \\
  &p\text{Clust}_{14} < 0.000024 \quad \text{AND} \\
  &p\text{Clust}_{15} < 0.000024 \quad \text{AND} \\

\end{align*}
\]
pCluster_4_3 > 0.014222 AND pCluster_4_0 <= 0.032689: MLM (28.0/1.0)
pCluster_4_3 <= 0.014803 AND pCluster_2_0 <= 0.03813 AND pCluster_3_0 <= 0.487121 AND pCluster_1_2 > 0.002153 AND pCluster_1_1 <= 0.000222: MlM (25.0)
pCluster_2_0 > 0.245194 AND pCluster_3_0 > 0.487121 AND pCluster_3_0 <= 0.487819: MMM (3.0)
pCluster_1_2 <= 0.000969 AND pCluster_1_0 <= 0.002891 AND pCluster_0_1 <= 0.011: MlM (159.0/29.0)
pCluster_3_3 > 0.005132 AND pCluster_1_1 <= 0.000126: MlH (3.0)
pCluster_3_3 <= 0.005132 AND pCluster_1_0 > 0.003551 AND pCluster_1_1 <= 0.000033 AND pCluster_3_2 <= 0.010386: MlM (7.0)
pCluster_0_1 <= 0.001067 AND pCluster_3_3 <= 0.002324 AND pCluster_1_3 <= 0.000021: MLM (4.0/1.0)
pCluster_3_3 <= 0.005132 AND pCluster_1_0 <= 0.004741 AND pCluster_0_1 <= 0.000568: MLM (2.0)
pCluster_3_3 <= 0.005132 AND pCluster_1_0 <= 0.004741 AND pCluster_1_1 <= 0.00268 AND pCluster_3_3 <= 0.002324 AND pCluster_1_1 > 0.000012 AND pCluster_1_0 <= 0.002296 AND pCluster_0_1 > 0.011592: MLM (12.0/1.0)

Rules generated by the PART techniques are as follow:

Village <= 81: MLH (640.0/4.0)
Village > 104 AND Village <= 241 AND Block <= 447 AND potash > 1 AND Village <= 154 AND Village > 152 AND Village <= 180: MlM (189.0/7.0)
Village > 181 AND Village <= 234 AND Village <= 212 AND Village > 190: MLM (84.0)
Village > 95 AND
srno > 95 AND
potash <= 1: MLM (56.0/1.0)  
Village > 185: MLM (32.0)  
srno > 95 AND  
Village <= 243 AND

Village <= 131 AND  
Village > 181 AND

potash <= 2 AND  
Village > 228 AND

Village > 86 AND  
Village <= 233: MLM (35.0)

Village > 95 AND  
Village <= 243 AND

Village <= 103 AND  
Village > 181 AND

Village <= 100: MLM (33.0)  
Village <= 217 AND

srno > 95 AND  
Village > 184 AND

Village <= 131 AND  
Village > 202 AND

potash <= 2 AND  
Block <= 447: MLM (25.0)

Village <= 86: MLM (31.0/1.0)  
Village <= 243 AND

Village > 104 AND  
Block <= 447 AND

Village <= 242 AND  
Village <= 98 AND

Block > 447 AND  
potash > 2: MLM (10.0)

phos <= 1 AND  
Village <= 243 AND

Village > 229 AND  
Village > 181 AND

Village > 234: MLM (57.0)  
Village <= 184: MLM (18.0)

Village > 104 AND  
Village <= 243 AND

Village <= 243 AND  
Village <= 98 AND

Block > 447 AND  
Block <= 1181: MLH (9.0/3.0)

Village <= 215 AND  
Village <= 243 AND

Village <= 214 AND  
Block <= 447 AND

Village <= 212: MLM (48.0)  
phos <= 1 AND

Village > 104 AND  
srno <= 148 AND

Village <= 243 AND  
srno <= 135: MLM (13.0/1.0)

Block > 447 AND  
Village <= 243 AND

Village > 215 AND  
Block <= 447 AND

Village > 218 AND  
phos <= 1 AND

Village <= 233 AND  
potash <= 2 AND

Village <= 227: MLM (68.0/3.0)  
Village > 187 AND

Village <= 131 AND  
Village > 196: MLL (8.0/1.0)

potash <= 2 AND  
Village <= 243 AND

Village > 87 AND  
Village <= 183 AND

Village > 90 AND  
phos <= 1 AND

Village > 101: MLM (44.0/6.0)  
potash >= 2 AND

Village <= 95 AND  
srno >= 747 AND

Village <= 94 AND  
srno >= 1062: MLL (13.0/4.0)

Village > 87 AND  
Village > 243: MLL (8.0/1.0)

Village <= 90: MLM (24.0)  
phos > 1 AND

srno > 91 AND  
Block > 447 AND

Village <= 95 AND  
Village > 215 AND

Village <= 94 AND  
srno > 580 AND

Village > 91: MLM (24.0)  
srno > 1071: MLM (3.0/1.0)

Village <= 95 AND  
phos > 1 AND

potash <= 2 AND  
Block <= 447: MLH (2.0/1.0)

srno <= 938: MLH (17.0)  
Block <= 447 AND

Village <= 243 AND  
potash <= 2 AND

Village > 181 AND  
srno <= 669 AND

Village <= 189 AND  
Village <= 183: MLL (11.0/1.0)
Village <= 215 AND
Village <= 214 AND
Block > 447 AND
Village > 213: MLM (8.0)
Village <= 215 AND
Block > 447 AND
srno <= 1426 AND
Village <= 214: MLM (6.0)
Village <= 215 AND
Block <= 447 AND
Village <= 214: MLM (15.0/6.0)

Block <= 447 AND
srno <= 1415: MLM (7.0)
Block > 447 AND
srno <= 796: MLM (3.0/1.0)
Village <= 215 AND
Block > 447: MMM (10.0/2.0)
Village <= 223 AND
srno <= 1415: MLM (7.0)
Block > 447 AND
Village > 238: MlM (8.0)
srno <= 460: MLM (7.0/1.0)
Block > 447 AND
phos <= 1 AND
Village <= 223 AND
srno <= 1567: MLL (5.0/1.0)
Village <= 223 AND
Block > 447 AND
Village <= 217: MLL (4.0)
Block > 447 AND
phos <= 1 AND
Village <= 231 AND
Village > 222: MMM (6.0/3.0)

Village <= 216: MLL (6.0/1.0)
: MLM (4.0/1.0)

The performance of the algorithm can be measured by using various parameters like TP rate, FP rate, recall, etc. True positive rate (TP rate) is the number of instance belongs to same class as specified by the algorithm divided by the total number of instance. False-positive rate (FP rate) is the number of instance doesn’t belong to the class specified by the algorithm divided by the total number of instances. Precision is the probability that randomly selected instance is correctly classified that can be given as

\[ \text{Precision} = \frac{TP}{TP + FP} \times 100\% \]

Recall is the average of probabilities of all instances within dataset.

\[ \text{Recall} = \frac{TP}{TP + FN} \times 100\% \]

F-measure is mean of precision and the recall can be given as

\[ \text{F-measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \]

Table 1 shows the comparison among PART, k-mean + Association and the proposed CCSA algorithm. The classification accuracy and no. of rules of the proposed algorithm is better than the other algorithms.. Overall a tree having compact rule and greater classification accuracy is generated by the proposed algorithm. Figure 2 and figure 3 shows the graphical comparison of accuracy as

---

**Classification Accuracy**

![Classification Accuracy Graph](image)

Fig. 2: Accuracy Comparison of tree by using various Algorithms PART, Kmean+ Association and proposed(CCSA) algorithm
Table 3: Comparison of various algorithms on Soil dataset

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Number of Rules</th>
<th>Classification accuracy</th>
<th>TP rate</th>
<th>FP rate</th>
<th>Precision</th>
<th>recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48(C4.5)</td>
<td>81</td>
<td>94.17</td>
<td>0.942</td>
<td>0.07</td>
<td>0.928</td>
<td>0.942</td>
<td>0.932</td>
</tr>
<tr>
<td>Kmean+J48</td>
<td>17</td>
<td>95.6</td>
<td>0.946</td>
<td>0.067</td>
<td>0.934</td>
<td>0.946</td>
<td>0.937</td>
</tr>
<tr>
<td>Schwarz based Kmean+J48</td>
<td>13</td>
<td>97.17</td>
<td>0.951</td>
<td>0.065</td>
<td>0.941</td>
<td>0.951</td>
<td>0.941</td>
</tr>
<tr>
<td>PART</td>
<td>40</td>
<td>94.69</td>
<td>0.947</td>
<td>0.047</td>
<td>0.944</td>
<td>0.947</td>
<td>0.945</td>
</tr>
<tr>
<td>Kmean+association</td>
<td>24</td>
<td>96.19</td>
<td>0.952</td>
<td>0.045</td>
<td>0.950</td>
<td>0.952</td>
<td>0.950</td>
</tr>
<tr>
<td>CCSA</td>
<td>13</td>
<td>97.69</td>
<td>0.956</td>
<td>0.042</td>
<td>0.955</td>
<td>0.956</td>
<td>0.955</td>
</tr>
</tbody>
</table>

well as other parameters for PART, k-mean + Association and the proposed CCSA algorithm. The algorithm can be compared on other datasets. Various datasets are downloaded that are available with WEKA are used to verify the performance of the proposed algorithm. The datasets used are diabetes and the ionosphere. Table 2 specifies various characteristics and performance comparison of the different algorithm on these datasets.

The table 3 shows the comparison of the various algorithms on the soil dataset. It can be seen no. of rules are reduced and classification accuracy as well as the TP rate of the proposed algorithm i.e. Schwarz based Kmean clustering cascaded by association based classification is higher than all the other techniques. The number of rules in the PART algorithm is 40 i.e. maximum while in the proposed algorithm is only 13. This shows the reduction in rules.

CONCLUSION

The paper produces an algorithm that results in reduced decision rules for classification. The algorithm is developed by cascading the clustering and association based classification algorithm. In the first step the SC (Schwarz Criterion) is applied to get the optimal number of clusters in the KMEAN clustering that is cascaded by association to get the decision tree. The algorithm is implemented using WEKA on the soil data set and two other datasets, and the result shows improved classification accuracy. The number of rules also compared to show the reduction in the rules. Various parameters like TP rate, FP rate, precision, recall, and f-measure are also evaluated to analyze the performance of the proposed algorithm. In future, the decision tree generating rules can be optimized and other recommendations can extended to other crops of different seasons.
REFERENCES