International Journal of Advanced Educational Research
ISSN: 2455-6157
Impact Factor: RJIF 5.12
www.educationjournal.org
Volume 3; Issue 1; January 2018; Page No. 438-441

Cascaded clustering and Ant-Miner based classification

Amit Kumar
Govt. PG College for Women, Rohtak, Haryana, India

Abstract
Data mining refers to extracting knowledge by analyzing data from different perspectives. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge driven decisions. The data is collected from different sources from different fields depending upon type of knowledge to be acquired. Classification is one of the basic data mining operations which aims to predict categorical data labels and helps in decision making. Ant-Miner is bio-inspired, ant colony based classification rules generator which is found much effective than other classifiers. Ant- Miner cannot be beneficial for a very large dataset. This paper can improve the Ant- Miner by using the concept of k-mean clustering, so that it can be applied on large data set. In this paper, data set "Weight Lifting Exercises monitored with Inertial Measurement Units Data Set" is used taken from "UCI repository". The simulation results show the better performance of modified Ant- Miner as compared to existing Ant- Miner with reduced tree size.

Keywords: data mining, ant-miner, k-mean, classification

1. Introduction
Data mining has proved to be an attractive field for the researchers, data analysts and knowledge workers. Data mining is concentrated on designing and employing computational methods to formulize a model based on a given knowledge representation from real-world structured data [1]. Knowledge discovery in databases (KDD) is the process of extracting models and hidden patterns from large databases. Data Mining (DM) refers to the process of applying the discovered algorithm to the data [2]. It is an important step of KDD. There are several data mining tasks like clustering and classification. Classification aims to predict unknown class labels. The popular category of classification model consists of classification rules that are expressed as IF-THEN rules. The aim of the classification algorithm is to discover a set of classification rules which can be further applied for decision making [3]. One algorithm for solving this task is Ant-Miner proposed by Parpinelli and colleagues [4], which employs ant colony optimization (ACO) techniques to discover classification rules of the form:
IF (term1 AND term2 AND ….. term) THEN (predicted class) where each term is of the form <attribute = value>. The number of terms in IF part can vary. The consequent of a rule is a predicted class when the given record satisfies all the terms in the rule antecedent. Classification rules have the advantage of representing knowledge at a high level of abstraction, so that they are intuitive to the user [5]. These rules are easy to interpret and have natural language representation.

2. Ant- Miner
Ant-Miner has produced good results when compared with more conventional data mining algorithms and it is still a relatively recent algorithm that suggests further research trying to improve it. In the original Ant-Miner, the goal of the algorithm was to produce an ordered list of rules, which was then applied to test data in the order in which they were discovered. This makes it difficult to interpret the rules at the end of the list, since their conditions make sense only in the context of all the previous rules in the ordered list of rules [5]. This makes the discovered rules easier for the user to interpret, now the interpretation of each rule is independent from all the other discovered rules.

A. The Ant-Miner Algorithm
Ant-Miner follows a sequential covering approach to discover a list of classification rules covering the maximum training cases. It starts with an empty list and the training set consists of all the training cases [4]. After every iteration, a rule is added to rule list and training set is reduced.
The original Ant-Miner algorithm upon which the rule set is generated is discussed here:

Input
Empty rule list, Training Set, Initial Pheromone trail, rule index, test index

Output
Discovered Rule List

Algorithm
1. Initialize training set to all training cases, an empty rule list and initialize all trails with the same amount of pheromone i.e.:
   - TrainingSet = {all training cases};
   - DiscoveredRuleList = [ ];
2. Initialize rule index and test index:
   - WHILE (TrainingSet > Max_uncovered_cases)
   - t = 1; /* ant index, and also rule index */

- j = 1; /* convergence test index */

3. Repeat
   - The ant starts with an empty rule and incrementally constructs a classification rule Rt by adding one term at a time to the current rule;
   - Prune rule Rt; /* remove irrelevant terms from rule */
   - Increase pheromone in the trail followed by Ant (proportional to the quality of Rt) and decrease pheromone in the other trails (simulating pheromone evaporation);
   - IF (Rt is equal to Rt – 1) /* update convergence test */
     THEN j = j + 1;
     ELSE j = 1;
     END IF
     t = t + 1;
   - UNTIL (i ≥ No_of_ants) OR (j ≥ No_rules_converg)
4. Select best rule Rbest among all rules Rt constructed by all the ants;
5. Add rule Rbest to DiscoveredRuleList; TrainingSet = TrainingSet - {set of cases correctly covered by Rbest}; End While

Ant-Miner discovers an ordered list of classification rules based on a heuristic function involving information gain – a popular heuristic function in data mining [3] - and positive feedback involving artificial pheromone. Individually iteration of the Repeat-Until loop an ant attempts to discover a rule by selecting terms in a probabilistic manner continuously all the attributes have been used to make the current rule or adding any other available term would make the rule coverage less than *min_cases_per_rule* – a user-specified threshold.
Discovered rule is then pruned in an attempt to reduce over fitting to the training data and increase rule quality. Later on, pheromone values for the terms in the current rule are increased in order to increase the probability that other ants will select those terms and then the pheromone values for all terms are normalized.
The While loop iterates until the number of training examples remaining in the dataset becomes less than or equal to *Max_uncovered_cases* – another user-specified threshold. The rule discovered in the Repeat-Until loop that has the highest quality is then added to the list of discovered rules and training examples correctly covered by that rule are removed from the training dataset [3].

B. K-Mean Clustering

K-means clustering is a partitioning based clustering technique of classifying/grouping items into k groups (where k is user specified number of clusters). The grouping is done by minimizing the sum of squared distances (Euclidean distances) between items and the corresponding centroid. A centroid (also called mean vector) is "the center of mass of a geometric object of uniform density". Although K-means is simple and can be used for a wide variety of data types, it is quite sensitive to initial positions of cluster centers [6]. There are two simple approaches to cluster center initialization i.e. either to select the initial values randomly, or to choose the first k samples of the data points. As an alternative, different sets of initial values are chosen (out of the data points) and the set, which is closest to optimal, is chosen. However, testing different initial sets is considered impracticable criteria, especially for large number of clusters, Ismail *et al* (1989). Therefore, different methods have been proposed in literature by Pena *et al*. (1999). Also, the computational complexity of original K-means algorithm is very high, especially for large data sets. Computer science has been widely adopted in different fields like agriculture. One reason is that an enormous amount of data has to be gathered and analyzed which is very hard or even impossible without making use of computer systems [6].
In data mining, k-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. This results in a partitioning of the data space into Voronoi cells. Given a set of observations (x1, x2, …, xn), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k sets (k ≤ n) S = {S1, S2, …, Sk} so as to minimize the within-cluster sum of squares (WCSS) [**7**]:

$$\arg\min \sum_{i=1}^{k} \sum_{x_{j \in s_i}} \left\| x_j - \mu_i \right\|^2 \qquad (1)$$

Where μi is the mean of points in Si.

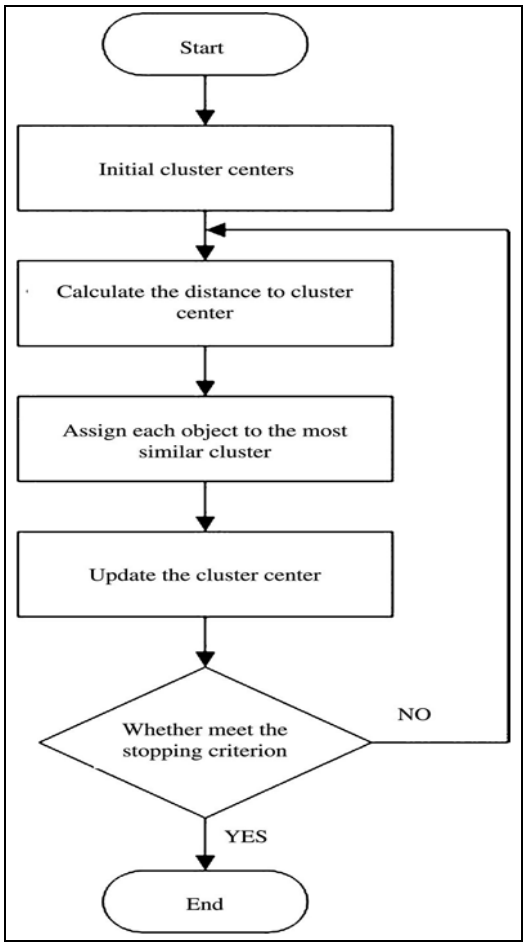

Fig 1: K Mean Clustering Process [8]

The basic step of k-means clustering is simple. In the beginning, we determine number of cluster K and we assume

the centroid or center of these clusters. We can take any random objects as the initial centroids or the first K objects can also serve as the initial centroids. Then the K means algorithm will do the four steps below until convergence [8]:

1. Determine the initial center coordinates.
2. Determine the distance of each object to the center.
3. Group the object based on minimum distance (find the closest centroid).
4. Update the cluster center until no object moves.

3. Proposed Algorithm

The proposed algorithm performs clustering then applies Ant-Miner on each cluster to classify the data. The basic idea is that Ant-Miner works well with small data set. So a large data set is broken into smaller chunks using clustering and then Ant-Miner is applied on it. Fig. 2 shows the step by step procedure of proposed algorithm.

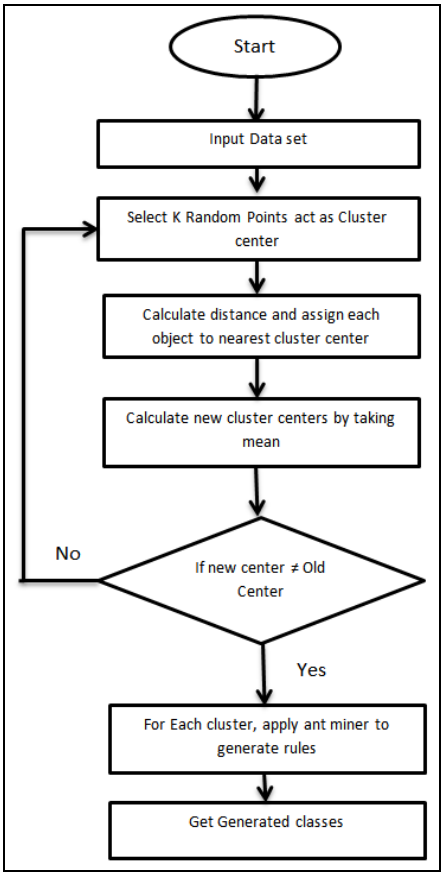

Fig 2: Flowchart of proposed algorithm

The algorithm can provide efficient results on large dataset. Large dataset is broken into small clusters by using general K-mean clustering algorithm. On each cluster, Ant Miner is applied separately that is capable of generating strong rules. Finally we get generated rules and their corresponding classes using novel cascaded approach.

4. Data set description

In this paper, "Weight Lifting Exercises monitored with Inertial Measurement Units Data Set" is used. This data set is taken from the website "UCI repository". By this data set, six young health subjects were asked to perform 5 variations of the biceps curl weight lifting exercise. One of the variations is the one predicted by the health professional [9].

Table 1: various characteristics of data set

Data Set Characteristics	Multivariate
Number of instances	5090
Number of attribute	59
Attribute characteristics	Real
Missing Values?	Yes

5. Results

The implementation of the proposed algorithm is performed using the WEKA tool on the dataset described in previous section. We have to add the new algorithm to the WEKA tool. New algorithm can be added to the WEKA by using the eclipse. The proposed algorithm and existing algorithm results are in following figures.

A. **Classification Accuracy and Size of Tree:** Fig 3 shows the comparison of existing approach with proposed approach in terms of classification accuracy and size of tree. It is clear from the results that proposed algorithm outperforms existing approach in given scenario.

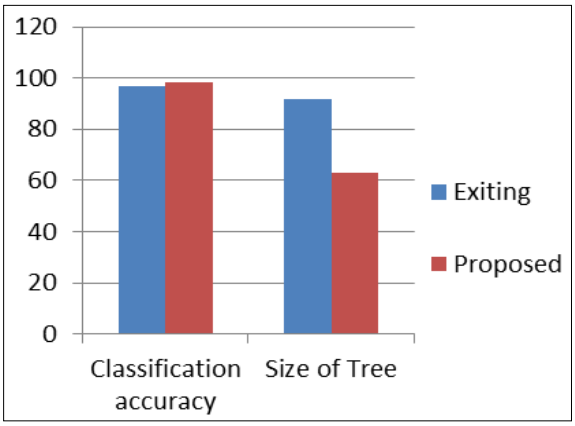

Fig 3: Comparison of classification accuracy and size of tree

B. **TP Rate:** In terms of TP rate, proposed algorithm is far better than existing algorithms as depicted by the results shown in Fig 4.

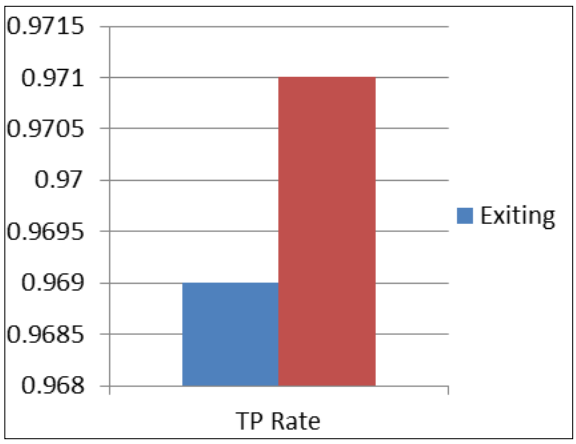

Fig 4: Comparison of existing and proposed algorithm on TP Rate

C. **Precision:** Results shown in Fig 5 show that proposed approach is better form precision point of view in given scenario when applied on large dataset.

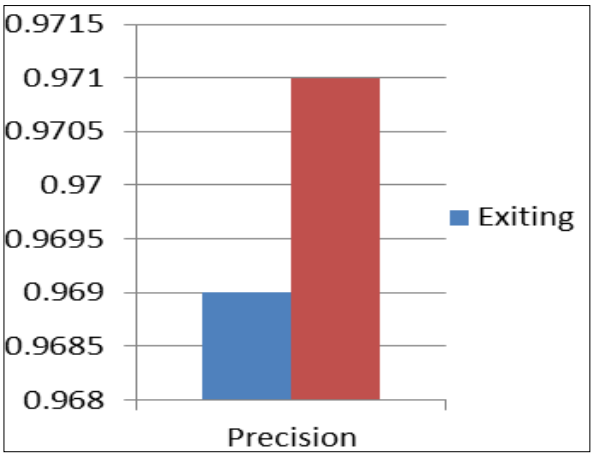

Fig 5: Comparison of existing and proposed algorithm on Precision

D. **Recall:** In Fig 6, comparison of recall of existing and proposed algorithms is given which shows that proposed algorithm outshines in this parameter also.

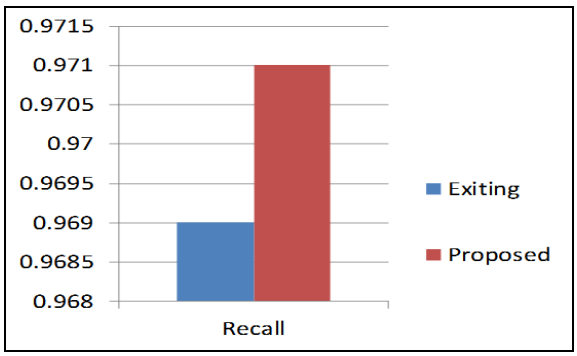

Fig 6: Comparison of existing and proposed algorithm on Recall

E. **FP Rate:** FP Rate of proposed algorithm reaches a scale of 0.011 as shown in Fig 7. Thus, proposed algorithm is better in this parameter too.

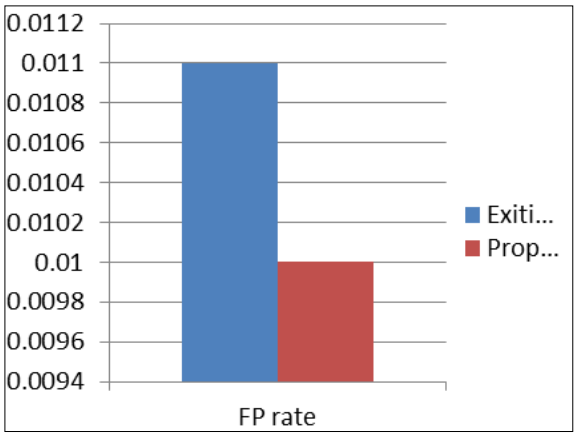

Fig 7: Comparison of existing and proposed algorithm on FP rate

6. Conclusion

Data mining is very effective technique to predict with the help of historical behaviour of data and statistics. ML (Machine learning) can often be successfully applied to these problems for improving the efficiency of systems and the designs of machines. As the Ant-Miner is efficient for small data sets, so we have partitioned data in to small set of clusters then apply the Ant miner for the classification. Specific parameters like recall, precision are optimized. The result shows that the modified Ant-Miner improves the accuracy approx. by 2%. It also reduces the tree size. Other parameters such as precision, TP rate, FP rate are compared for existing and proposed approaches. The results show that proposed algorithm is better in all said parameters. In future other swarm intelligence techniques can be applied for classification. The proposed algorithm results can be verified on other datasets.

7. References

1. Fernando EB, Otero *et al*. A New Sequential Covering Strategy for Inducing Classification Rules With Ant Colony Algorithms, IEEE Transactions on Evolutionary Computation. 2013; 17(1):64-76.
2. Liu B, Abbass HA, McKay B. Classification rule discovery with ant colony optimization. In Intelligent Agent Technology, IEEE/WIC/ACM International Conference on. IEEE Computer Society, 2003, 83-83.
3. James Smaldon *et al*. A New Version of the Ant-Miner Algorithm Discovering Unordered Rule Sets, GECCO'06, Seattle, Washington, USA. Copyright 2006 ACM 1-59593-186-4/06/0007…$5.00, 2006, 8-12.
4. Parpinelli RS, Lopes HS, Freitas AA. Data mining with an ant colony optimization algorithm. IEEE Transactions on Evolutionary Computing. 2002; 6(4):321-332.
5. Witten IH, Frank E. Data Mining: practical machine learning tools and techniques. 2nd Edition. Morgan Kaufmann, 2005.
6. Ritu Sharma Sachdeva. K-Means Clustering in Spatial Data Mining using Weka Interface, International Conference on Advances in Communication and Computing Technologies (ICACACT) Proceedings published by International Journal of Computer Applications (IJCA), 2012.
7. Namita Bhan. Comparative Study of EM and K-Means Clustering Techniques In WEKA Inter-Face, International Journal of Advanced Technology & Engineering Research (IJATER), 2013; 3(4).
8. Sapna Jain. K-MEANS Clustering Using WEKA Interface, Proceedings of the 4th National Conference; INDIACom-2010 Computing For Nation Development, 2010.
9. Velloso E, Bulling A, Gellersen H, Ugulino W, Fuks H. Qualitative Activity Recognition of Weight Lifting Exercises. Proceedings of 4th International Conference in Cooperation with SIGCHI (Augmented Human '13). Stuttgart, Germany: ACM SIGCHI, 2013.