

Correlation Associative Rule Induction Algorithm Using ACO

C. Nalini

Department of Information Technology, Kongu Engineering College, Erode, India
Email: nalinikec@gmail.com

Received 2 May 2016; accepted 12 May 2016; published 4 August 2016

Copyright © 2016 by author and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY).
<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Classification and association rule mining are used to take decisions based on relationships between attributes and help decision makers to take correct decisions at right time. Associative classification first generates class based association rules and use that generate rule set which is used to predict the class label for unseen data. The large data sets may have many null-transactions. A null-transaction is a transaction that does not contain any of the itemsets being examined. It is important to consider the null invariance property when selecting appropriate interesting measures in the correlation analysis. Real time data set has mixed attributes. Analyze the mixed attribute data set is not easy. Hence, the proposed work uses cosine measure to avoid the influence of null transactions during rule generation. It employs mixed-kernel probability density function (PDF) to handle continuous attributes during data analysis. It has ably to handle both nominal and continuous attributes and generates mixed attribute rule set. To explore the search space efficiently it applies Ant Colony Optimization (ACO). The public data sets are used to analyze the performance of the algorithm. The results illustrate that the support-confidence framework with a correlation measure generates more accurate simple rule set and discover more interesting rules.

Keywords

Associative Classification, Mixed Data, Correlation, ACO, Mixed-Kernel PDF

1. Introduction

Frequent pattern mining generates frequent patterns in a given dataset. These patterns show interesting relationships among attribute-value pairs which occur frequently during frequent pattern mining. Association rules are generated from the frequent patterns and such rules are used in the decision-making process. An associative classification algorithm generates classification rules from class-based association rules. It is a two-step process

which consists of frequent pattern mining followed by rule generation. There are many associative classification algorithms available like Classification Based on Association (CBA), Classification Based on Multiple Association Rules (CMAR), and Classification Based on Predictive Association Rules (CPAR), but they differ in their approaches used for frequent itemset mining and in the way the discovered rules are analyzed and used for classification. Associative classification is an alternative classification method for building rules based on frequent patterns in a dataset. The main difference between association rule mining and classification is the outcome of the rules generated. The search for rules in classification is directed towards the class attribute, whereas the search for association rules is not directed towards any specific attribute. In associative classification, class-based association rules are generated during association rule mining referred to as the Class Association Rules (CARs). Then, Classification Rule Mining (CRM) is applied on the CARs to generate a Classifier model or rule set. The Associative classification algorithm is used in various applications like customer behavior prediction, portfolio risk management, Intrusion and detection systems etc.

Swarm Intelligence (SI) is a population-based computational technique. It has been used to solve problems in artificial intelligence and machine learning. The characteristics of swarm intelligence techniques are reliability, convergence capability and parallelism, attracts many researchers. Marco Dorigo has been introduced Ant Colony Optimization (ACO) in 1992 inspired by ants behavior. In nature, Ants use Pheromone to pass information among them. The shortest path to the food source is selected based on the thickness of pheromone in a path. In ACO, an artificial ant adds solutions in sequential order to the solution space based on the probability of pheromone and heuristic function. The ants cooperate with one another to find good-quality solutions for problems in large search spaces. ACO use heuristic search to add solutions to the solution space.

In association rule mining, support measure is used to measure the usefulness of a rule, (*i.e.*) the ratio of transactions having an itemset appears with respect to the total number of transactions. And confidence is used to measure certainty of the discovered rule (*i.e.*) the number of transactions that contain all the items in the consequent as well as the antecedent to the number of transactions that contain all items in the antecedent.

$$\text{Support}(X \Rightarrow Y) = P(X \cup Y) \quad (1)$$

$$\text{Confidence}(X \Rightarrow Y) = P(Y/X) \quad (2)$$

Sometimes, support and confidence are not sufficient to discover interesting rule from the data set. In addition to, Correlation measures are used to find an interesting pattern from the data set.

Classification problem is a combinatorial optimization problem. Combinatorial optimization is the process of finding one or more optimal solutions in a well-defined discrete problem space. In associative classification, rule generation is based on frequent itemset mining processes so it takes more time for large databases. A decision tree algorithm Iterative Dichotomizer (ID3) was proposed by Quinlan in [1]. Later, Quinlan [1] extended this algorithm and named it as C4.5, which became a benchmark algorithm for comparing other classification algorithms. It adopts a top-down greedy approach to generate decision trees. To solve over fit problems, it starts with a set of training datasets and build a decision tree during the learning process. To build a tree it partitions the dataset recursively into smaller subsets. It employs information gain as an attribute selection measure to find the best split point. The attribute with the highest information gain is chosen as the split attribute or the split point. The computational complexity of the algorithm is $O(n*|D|\log(|D|))$, where n is the number of attributes describing the tuple and $|D|$ is the number of tuple in the training dataset.

One of the earliest and simplest association-based classification algorithms is Classification Based on Association (CBA). It uses the iteration-based apriori algorithm for frequent pattern mining and the strong class-based association rules are used to create a classification model. It employs heuristic search to build classifiers, where rule-based ordering is used to solve the conflict rule strategy. It was empirically found to be more accurate than C4.5 on a good number of datasets. Classification Based on Multiple Association Rules (CMAR) [2] adopts FP-growth algorithm to generate a frequent pattern. In rule-based classification, if a rule is satisfied by a tuple X , then the rule is said to be triggered. If more than one rule is triggered then a conflict resolution strategy method is invoked to solve the problem. To correct this problem, CMAR divides the rules into groups according to class labels and uses a weighted chi-squared measure to find the strongly correlated rules within a group, generate a more accurate and efficient rule set than CBA.

CBA and CMAR adopt minimum support and confidence framework to implement frequent pattern mining. However, it generates large rule set. To reduce this, Classification Based on Predictive Association Rules (CPAR)

algorithm [3] applies FOIL measure to prune the rule set. It is used to solve multi-class problems. Similar to CMAR, it groups rules according to the class label but selects the best k rules of each group to predict the class label of a tuple X. This procedure evades the influence of low-ranked rules. Compared to CMAR, it generates a more accurate and efficient rule set.

ACO is a population-based meta-heuristic algorithm used for solving combinatorial optimization problems. When choosing a method to tackle a given combinatorial optimization problem, several factors must be taken into consideration like the time taken to deliver the results, the quality of the solution found, the quality of the results and its ability to scale up or down with respect to the problem size.

The first ACO-based classification rule discovery algorithm called Ant Miner proposed by Parpinelli *et al.* [4]. It uses sequential rule induction algorithm to generate a rule set. Ant starts with an empty solution set and incrementally adds the terms one by one at a time based on the heuristic quality of the term (information gain) and the amount of pheromone deposited on it. It generates ordered classification rule set. It discretizes the numerical attribute values into nominal.

Afterwards, Liu *et al.* have been extended Ant Miner into two versions such as Ant Miner2 [5] and Ant Miner3 [6]. Ant Miner2 make use of density estimation as a heuristic function. It has been produced the same result with minimum time complexity. Ant Miner3 update and evaporate the pheromone only those terms occurred in the rule instead of all the terms. It has been applied pseudo-random proportional transition rule to support both exploitation and exploration. It generates more accurate rule set than Ant Miner and Ant Miner2 on a good number of datasets. Both the algorithms support only categorical attributes. J. Smaldon *et al.* [7] extended Ant Miner to discover unordered rule set from a dataset. If a test tuple X is satisfied by more than two rules, then the higher rule quality-class label is assigned to X. The results illustrated that Ant Miner performs well with the unordered approach.

In general, ACO supports only nominal attributes. Socha [8] extended ACO to a continuous domain, used Mixed-Kernel PDF to represent pheromone for a continuous attribute. Nalini *et al.* proposed discovering unordered rule sets for mixed variables using an ant-miner algorithm in [9]. A hybrid ACO/PSO algorithm proposed in [10], supports mixed attributes which apply PSO algorithm to handle continuous valued attributes and apply ACO algorithm to handle nominal attributes. It generates mixed attribute rule set. Baig *et al.* proposed [11] a new ACO based classification algorithm called Ant Miner-C. It implements a correlation-based heuristic function in order to guide the ant to select items that help to discover a rule.

The existing associative classification algorithms apply only support and confidence frame work to evaluate the interesting pattern. The support and confidence measures generate a large number of association rules, most of which are not interesting to the users. The large data sets may have many null-transactions. It is important to consider the null invariance property when selecting appropriate interesting measures in the correlation analysis. The objective of the proposed work is to generate correlation based rule set. It uses cosine measure to avoid the influence of null transactions during rule generation. The proposed work generates correlation based associative rule set. ACO is used to generate optimized associative classification rule set for mixed data. To improve the accuracy of the rule set and reduce the number of redundancies it applies correlation based heuristic function.

2. The Proposed Algorithm

The proposed correlation based associative rule induction algorithm employ ACO to generate optimized associative classifier model. Most association rule mining algorithms employ a support and confidence framework, help to weed out uninteresting rules. But many of the rules are not interesting to the users. To overcome this problem the proposed work employs cosine measure in addition to support and confidence threshold to evaluate the generating pattern. It helps the ant to generate optimized rule in each generation. After each generation, the pheromone values are updated by using suitable pheromone update function. Mixed-kernel PDF function is used to initialize and update the pheromone level for continuous attributes. The mixed-kernel PDF is calculated by using the Equation (3)

$$P(x) = G(x, \omega, \mu, \sigma) = \sum_{j=1}^k \omega_j \cdot g(x, \mu_j, \sigma_j) \quad (3)$$

where,

ω = vector of weights associated with the mixture, μ =vector of means, σ =vector of standard deviations.

k = no of ranges.

An ant starts with an empty rule and incrementally adds terms one by one. The selection of a term to be added is probabilistic. It is based on two factors: a heuristic quality of the term and the amount of pheromone deposited on it by the previous ants.

$$P_{ik} = \frac{\tau_{ij}^{\alpha}(t) \eta_{ij}^{\beta}(s)}{\sum_{i=1}^a x_i \sum_{j=1}^{b_i} \tau_{ij}^{\alpha}(t) \eta_{ij}^{\beta}(s)} \quad (4)$$

where $\tau_{ij}(t)$ is the amount of pheromone associated with the edge between term_i and term_j for the current ant, it may change after the passage of each ant, $\eta_{ij}(s)$ is the value of the heuristic function for the current iteration s. a is the total number of attributes, x_i is a binary variable that is set to 1 if the attribute a_i has not been used by the current ant and else set to 0, and b_i is the number of values in the a_i attribute's domain. The denominator is used to normalize the numerator value for all the possible choices. The parameters alpha (α) and beta (β) are used to control the relative importance of the pheromones and heuristic values in the probability determination of the next movement of the ant. Here assign $\alpha = \beta = 1$, which means that give equal importance to the pheromone and heuristic values. **Figure 1** depicts the steps involved in the proposed algorithm.

The heuristic function guides an ant to move forward in the search space. The algorithm use correlation based heuristic function to choose the term (attribute–value pair). The heuristic value is calculated by using the Equation (5).

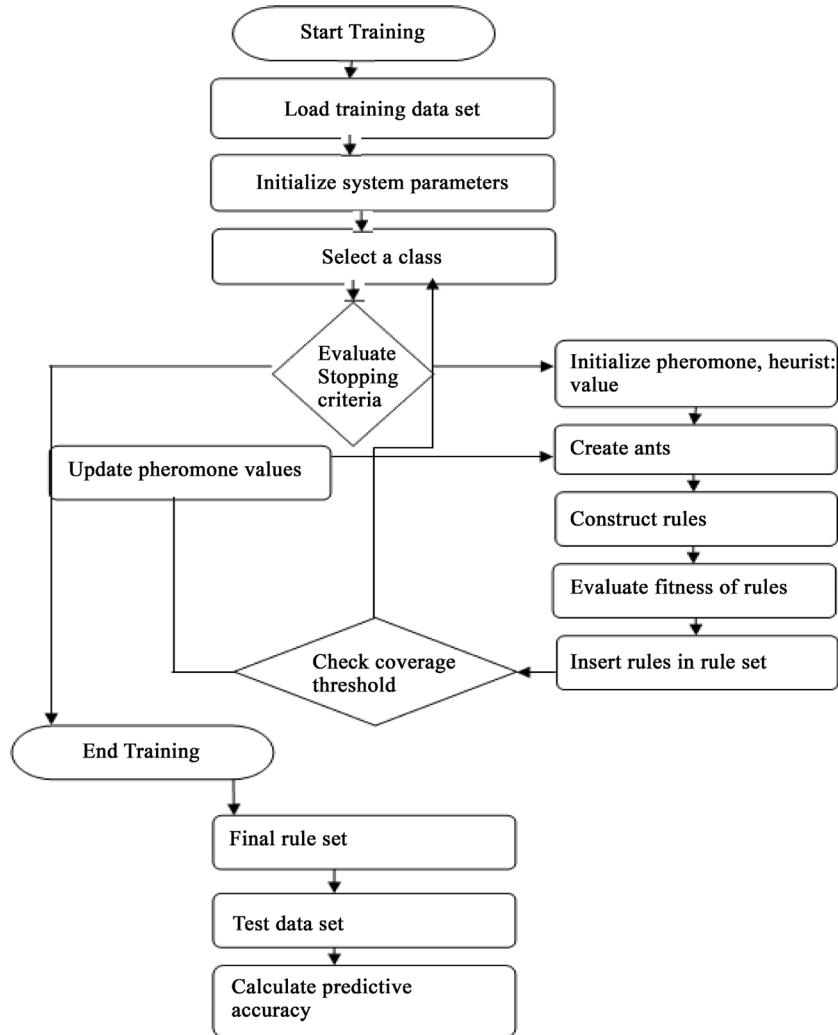


Figure 1. Flow chart of the proposed algorithm.

$$\eta_{ij} = P(term_j / term_i, class_k) \cdot P(class_k / term_j) \quad (5)$$

The correlation-based heuristic function find relation between the term to be added to most recently added term and the class label. It helps to explore the high dimensional search space and generate an efficient rule set. The quality (Q) of a rule is computed by using F-Score. It is the harmonic mean of precision and recall. It gives equal weight to precision and recall. Precision specify the exactness and recall denote the completeness of a rule.

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (6)$$

$$\text{recall} = \frac{\text{TP}}{\text{P}} \quad (7)$$

$$\text{F Score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (8)$$

In each iteration the best rule is stored and used to update the pheromone levels of the terms before the next iteration is initiated. In the case of nominal attributes, the pheromone level is increased using the Equation (9).

$$\tau_{ij}(t+1) = \tau_{ij}(t) + \tau_{ij}(t) \cdot Q \quad (9)$$

where:

$\tau_{ij}(t)$ is the pheromone level of the term $_{ij}$ at iteration t ,

Q is the quality of the rule constructed.

For continuous attributes, a new normal kernel is added to the mixed-kernel PDF. The mean and the standard deviation of the added kernel are the mean and the standard deviations of the range that was selected by the ant during the iteration. The addition of a new kernel changes the shape of the mixed-kernel PDF curve and thus changes the values of the pheromones for the neighboring ranges. The impact of this increase in pheromone on the pheromone value of a neighboring range decreases as the distance between the means of the added range and the neighboring range increases.

3. Result Analysis

The proposed algorithm is implemented by using C#.NET and conducted experiments on a machine that has 1.75 GHz dual processors with 1 GB RAM. We compared the results of the proposed approach with other well-known classification algorithms like Ant Miner, Ant Miner-C, and C4.5 based on the factors like the predictive accuracy, the number of rules generated and the number of terms per rule generated. The experiments are performed using a ten-fold cross-validation. The results of the ten runs are averaged, and this average is reported as the final result. The proposed work used public data sets (UCI repository) [12] to analyze the performance of the algorithm. **Table 1 & Table 2** portray the value of user defined parameters and the data sets used to analyze the performance of the algorithms.

Tables 3-5 illustrate the performance of the correlation associative rule induction algorithm using ACO against existing algorithms and **Figure 2** demonstrates the predictive accuracy of the proposed algorithm. The proposed algorithm applies cosine correlation measure to evaluate the interesting pattern. The null-invariant property of cosine measure helps to find interesting correlation based frequent pattern in most of the dataset, helps to enhance the predictive accuracy of associative classifier model. The results illustrate that the algorithm

Table 1. Parameter setting.

Parameter	Value
Number of Ants	1000
Evaporation Rate	0.15
Alpha, Beta	1
Minimum Support	0.02
Minimum Confidence	0.5
Minimum Coverage	0.98

Table 2. Datasets.

Datasets	No. of samples	No. of attributes	Attribute type	No. of classes
Zoo	101	18	Mixed	7
Wine	178	13	Numeric	3
Car	1728	6	Categorical	4
Glass	214	10	Numeric	7
Iris	150	4	Numeric	3
Credit-Australia	690	15	Mixed	2
Tic_tac_toe	958	9	Categorical	2
Ecoli	336	8	Numeric	8

Table 3. Average predictive accuracy.

Dataset	Proposed	Ant Miner-C	Ant Miner	C4.5
Zoo	98.0 ± 3.83	96.0 ± 5.16	81.36 ± 11.30	94.0 ± 9.17
Wine	99.0 ± 2.13	99.44 ± 1.76	90.0 ± 9.22	96.60 ± 3.93
Car	98.43+1.17	98.03 ± 1.17	82.38 ± 2.42	96.0 ± 2.13
Glass	95.50 ± 4.01	82.27 ± 6.67	53.33 ± 4.38	68.9 ± 8.98
Iris	99.87 ± 2.81	99.8 ± 4.50	95.33 ± 4.5	94.0 ± 6.63
Credit (Aust)	98.84 ± 1.14	82.44 ± 7.31	85.77 ± 4.75	79.37 ± 4.57
Tic-Tac-Toe	99.06 ± 0.92	100 ± 0.0	74.95 ± 4.26	94.03 ± 2.44
Ecoli	97.80 ± 1.48	83.64 ± 6.11	47.52 ± 11.32	82.99 ± 7.72

Table 4. Average number of rules generated.

Dataset	Proposed	Ant Miner-C	Ant Miner	C4.5
Zoo	5.00	7.00	5.10	7.0
Wine	5.10	6.00	5.50	5.30
Car	11.5	5.80	11.40	80.26
Glass	18.90	42.20	15.50	15.40
Iris	8.89	12.80	9.20	5.50
Credit (Aust)	6.80	5.50	3.90	74.80
Tic-Tac-Toe	11.20	12.30	6.60	38.60
Ecoli	15.50	57.60	8.60	14.00

Table 5. Average number of terms/rules generated.

Data set	Proposed	Ant Miner-C	Ant Miner	C4.5
Zoo	1.80	7.00	5.10	7.60
Wine	1.00	1.25	1.04	1.41
Car	2.48	2.5	1.03	2.59
Glass	1.57	1.93	1.01	2.83
Iris	1.09	1.05	1.00	1.22
Credit (Aust)	1.35	1.57	1.00	3.22
Tic-Tac-Toe	1.72	2.74	1.09	2.64
Ecoli	1.67	1.75	1.01	2.84

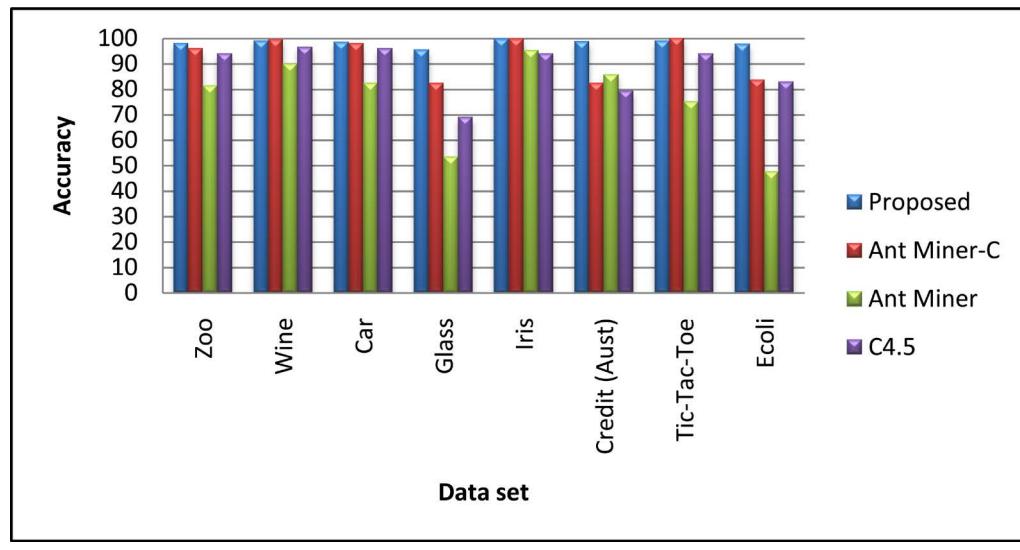


Figure 2. Analysis of predictive accuracy.

produced more accurate simple rule set than the existing algorithm. The existing algorithms use only support and confidence framework to evaluate the frequent pattern. The support and confidence measures generate a large number of association rules, most of which are not interesting to the users. From the results it concludes that the support-confidence framework with a correlation measure generates more accurate simple correlation rule set and discovers more interesting rules. It concluded the algorithm is good for high-dimensional large data set.

4. Conclusion

Data classification plays an important role in the data mining. It builds a classification model to classify new data. Association rules show the strong associations between attribute-value pairs that occur frequently in the data set. Associative classification integrates both association rule mining and classification to build more accurate classifier model than traditional classification algorithm like decision tree and rule induction algorithm. The proposed work generates correlation based associative rule set. The correlation based associative rule induction algorithm employed cosine measures to correlation analysis to avoid the influence of null transactions during rule generation. The null-invariant property of cosine measure helped to generate more accurate and simple rule set for mixed data. It constructs a competent and effective decision model to classify large data set. This algorithm is suitable to solve real time problems. In future, feature selection may be incorporated to select task-relevant features which may reduce the execution time.

References

- [1] Quinlan, J.R. (1996) Improved Use of Continuous Attributes in C4.5. *Journal of Artificial Intelligence Research*, **4**, 77-90.
- [2] Li, W., Han, J. and Pei, J. (2001) CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules. *Proceedings of IEEE International Conference on Data Mining*, 369-376.
- [3] Yin, X. and Han, J. (2003) CPAR: Classification Based on Predictive Association Rule. *Proceedings of the SIAM International Conference on Data Mining*, SIAM Press, San Francisco, 369-376.
<http://dx.doi.org/10.1137/1.9781611972733.40>
- [4] Parpinelli, R., Lopes, H. and Freitas, A. (2002) A Data Mining with an Ant Colony Optimization Algorithm. *IEEE Transactions on Evolutionary Computing*, **6**, 321-332. <http://dx.doi.org/10.1109/TEVC.2002.802452>
- [5] Li, W., Han, J. and Pei, J. (2001) CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules. *Proceedings of IEEE International Conference on Data Mining*, 369-376.
- [6] Liu, B., Abbass, H.A. and McKay, B. (2002) Density-Based Heuristic for Rule Discovery with Ant-Miner. *Proceedings of 6th Australia-Japan Joint Workshop on Intelligent Evolutionary Systems*, Canberra, 180-184.

- [7] Smaldon, J. and Freitas, A. (2006) A New Version of the Ant-Miner Algorithm Discovering Unordered Rule Sets. *GECCO'06*, Seattle, 43-50. <http://dx.doi.org/10.1145/1143997.1144004>
- [8] Socha, K. (2004) ACO for Continuous and Mixed-Variable Optimization. *Proceedings of the Fourth International Workshop on Ant Colony Optimization and Swarm Intelligence*, Brussels. http://dx.doi.org/10.1007/978-3-540-28646-2_3
- [9] Nalini, C. and Balasubramanie, P. (2008) Discovering Unordered Rule Sets for Mixed Variables Using an Ant-Miner Algorithm. *Data Science Journal*, **7**, 76-87. <http://dx.doi.org/10.2481/dsj.7.76>
- [10] Holden, N. and Freitas, A.A. (2008) A Hybrid PSO/ACO Algorithm for Discovering Classification Rules in Data Mining. *Journal of Artificial Evolution and Applications*, 1-11. <http://dx.doi.org/10.1155/2008/316145>
- [11] Baig, A.R. and Shahzad, W. (2012) A Correlation-Based ant Miner for Classification Rule Discovery. *Neural Computing and Applications*, **21**, 219-235. <http://dx.doi.org/10.1007/s00521-010-0490-5>
- [12] Merz, C. and Murphy, P. (1996) UCI Repository of Machine Learning Databases. University of California, Department of Information and Computer Science, Irvine.



Scientific Research Publishing

Submit or recommend next manuscript to SCIRP and we will provide best service for you:

Accepting pre-submission inquiries through Email, Facebook, LinkedIn, Twitter, etc.

A wide selection of journals (inclusive of 9 subjects, more than 200 journals)

Providing 24-hour high-quality service

User-friendly online submission system

Fair and swift peer-review system

Efficient typesetting and proofreading procedure

Display of the result of downloads and visits, as well as the number of cited articles

Maximum dissemination of your research work

Submit your manuscript at: <http://paperSubmission.scirp.org/>