
Protein classification via an ant-inspired association rules-based classifier

Muhammad Asif Khan* and Waseem Shahzad

Department of Computer Science,
National University of Computer and Emerging Sciences (NUCES),
Sector H-11/4, Islamabad, Pakistan
E-mail: masif.khan@nu.edu.pk
E-mail: waseem.shahzad@nu.edu.pk
*Corresponding author

Abdul Rauf Baig

Department of Computer Science,
National University of Computer and Emerging Sciences (NUCES),
Sector H-11/4, Islamabad, Pakistan
and
College of Computer and Information Sciences,
Al Imam Mohammad Ibn Saud Islamic University (IMSIU),
P.O. Box 5701, Riyadh 11432, Saudi Arabia
E-mail: raufbaig@ccis.imamu.edu.sa
E-mail: rauf.baig@nu.edu.pk

Abstract: Association rules mining and classification rules discovery are two important data mining techniques used to expose the relations among large sets of data items. The technique aims to find out the rules that satisfy the predefined minimum support and the confidence. Association rules mining has successfully been implemented in biomedical research and has demonstrated encouraging results in analysing the gene expression data in order to discover the relevant biological association among different genes, gene expression, and various protein properties like protein functionality and sequence similarity. In this paper, we applied the association rule mining technique – the ACO-AC to the problem of classifying proteins into its correct fold of the SCOP dataset. The technique combines the association rules mining and supervised classification mechanism using ant colony optimisation. Experimental results reveal the classifier performance in protein classification problem as excellent by identifying most accurate and compact rules.

Keywords: association rules mining; classification; rules discovery; structural classification of proteins; SCOP; ant colony optimisation; ACO.

Reference to this paper should be made as follows: Khan, M.A., Shahzad, W. and Baig, A.R. (xxxx) 'Protein classification via an ant-inspired association rules-based classifier', *Int. J. Bio-Inspired Computation*, Vol. X, No. Y, pp.xxx-xxx.

Biographical notes: Muhammad Asif Khan is a PhD student at NUCES, Islamabad, Pakistan. He received his BSc in Computer Science from Islamia College Peshawar, University of Peshawar. His research area includes protein classification, domain homology prediction and data mining. He has also an association with Orengo Research Group at the University College London (UCL), UK.

Waseem Shahzad received his PhD at NUCES Islamabad, Pakistan and is currently working as an Assistant Professor in the same university. He has research interests in the fields of data mining, computational intelligence, machine learning, theory of computation and soft computing. He has several research publications in these areas.

Abdul Rauf Baig is a Professor at Al Imam Mohammad Ibn Saud Islamic University (IMSIU), Saudi Arabia. He is also associated with NUCES, Islamabad, Pakistan. He received his PhD in Computer Science from University of Rennes-1, France. His research interests are in the areas of computational intelligence and machine learning. He has several research articles published in international journals and has completed many R&D projects. He is also a reviewer for several reputed international journals.

1 Introduction

The highly sophisticated methods applied on functional genomics have resulted to produce a huge amount of genomic data. The number of known protein sequences growing at very high rate are far larger than the number of known proteins structures. Despite recent technological advancement and structural determination methods, it is still difficult to find the structures of hundreds of protein sequences. Protein has different structures – the primary structure, secondary structure, tertiary structure and quaternary structure. A protein in tertiary structure is said to be a functional protein if it is folded correctly. Protein folding is the process through which protein primary structure transforms into its tertiary structure. During protein folding process, amino acids interact with other amino acids to form a well-defined three-dimensional structure of a protein also called the native state of a protein. Folding may be either successful or not. A successful folding is necessary for a protein to function properly. For protein fold recognition problem the terms classification and prediction are often used interchangeably. Classifying the three-dimensional protein structure is the major target of computational biology (Dill et al., 2008).

Getting maximum benefit of the voluminous genomic data for the betterment of humanity is vital. Data mining approaches attracted the world to deal with such kind of big data to extract useful information. Data mining is a combination of techniques that dig out novel, meaningful and valuable information using large databases and is very much useful for different tasks including classification, clustering and regression analysis. Classification is one of the important data analysis tasks that group items based on similar hidden characteristics matched with a set of already known labelled data items (Giannopoulou, 2008). This reveals the classification as a supervised technique having data items of different classes or groups with known class labels in advance. Classification builds a set of models based on training dataset and are used for testing dataset to correctly classify the class of a query data item. Training dataset consists of data samples representing different characteristics and their class labels. Among various models build by classification algorithm, an optimised model is used to classify a query data item of the testing dataset.

Mathematica routines were developed with artificial neural network for predicting the amino acids spatial proximity in order to view the relationship between primary structure and three-dimensional protein structure (Fairchild et al., 1995). Using amino acid frequencies to recognise the fold (Du et al., 2003; Taguchi and Gromiha, 2007), the PSMACA (Sree et al., 2013), SARAMA (Basu et al., 2013) are the few among those that investigate the relationship between residues and protein structure. Differential evolution (DE) strategy using HP model (Bitello and Lopes, 2006), lattice model (Hart and Newman, 2005), neural network (Igel et al., 2004; Langlois et al., 2004), gating neural network (Huang et al., 2003), support vector machines (Ding and Dubchak, 2001; Langlois et al., 2006), and K-local hyperplane distance nearest neighbour

(Okun, 2004) have also made their contribution in solving the protein classification problem. Sometimes, the three-dimensional structure information of a protein is missing. Despite the availability of its amino acids sequence it is very hard to obtain the 3-D coordinates of protein atoms. Such situation was targeted by Kin et al. (2004) and proposed the use of kernel matrix that includes kernel values representing the protein 3-D structure information and the missing entries. Moreover, the hierarchical strategy for protein and structure classification (Cheng and Baldi, 2006; Marsolo et al., 2005), ensemble classifier (Shen et al., 2006), combinatorial fusion technique (Lin et al., 2007), hidden Markov models, NN, SVM, Bayesian methods and clustering techniques (Cheng et al., 2008), protein secondary structures (Liu and Wang 2007) are the efforts toward solving the protein folding problem. In addition, among various evolutionary technique a GA-based and feature selection approach (Chen et al., 2009), ACO-based with 2-D HP model (Hu et al., 2008), HP model in a 3D cubic lattice model (Fidanova, 2006) have also played their role for tackling the problem.

Due to the speedy growth in genomic data in recent years, researchers are more concern about how to attain new information, knowledge and ideas of such voluminous biological data. Traditional methods lack the ability to meet this challenge. Association rule mining is one of the best choice of most of the researchers that extract the relevant and vital association between various gene expressions, protein sequences and structures. One such an approach was introduced by Yang et al. (2010) to discover the rules in order to predict protein secondary structure using support vector machines and the knowledge discovery process. Association rule mining concept was first introduced by Agrawal et al. (1993). The goal of the technique is to extract meaningful information in the form of correlations or associations among sets of data items of large databases.

In biomedical research, association rule mining discovers the rules for relevant information like patient symptoms, diagnosis and patient treatment procedure (Doddi et al., 2001). In genomic data analysis, association rule mining has successfully been applied to gene expression data analysis, protein-protein interactions, protein function and sequence motifs (Kotlyar and Jurisica, 2006). Solving the protein folding problem, a number of optimisation algorithms like evolutionary algorithm, ant colony optimisation (ACO) algorithm and Monte Carlo methods have been used. Association rule mining in association with ACO algorithm was used to mine those association rules that meet the pre-defined criteria and covering the whole data used during the training. From the last couple of decades, the algorithm attracted many researchers because of its successful contribution in various application areas and more importantly solving the optimisation problems (Grosan et al., 2006).

In this work, we explored the ACO-based association rule mining techniques known as the hybrid classification algorithm ACO-AC (Shahzad and Baig, 2011) for classification of proteins into its various folds of the

structural classification of proteins (SCOP) dataset. The method is the merger of the association rules mining and supervised classification algorithm – the ACO. Using the evolutionary characteristics, the algorithm searches the most appropriate and efficient set of association rules. The searching process involved a pre-defined minimum support and a confidence threshold for rules that build a classifier. Rules discovery in this approach is an iterative process during which efficient rules are selected, build a prediction model and hence finally adopted for classification of unseen samples.

2 Material and methods

2.1 SCOP dataset

The dataset used in this work is the most well known and authentic protein database – the SCOP. It provides the structural as well as evolutionary relationships among all proteins of known structures in a comprehensive manner (Hubbard et al., 1997). Considerable amounts of human efforts have been involved in classification of proteins through visual inspection and structures evaluation.

The protein classification in SCOP is on hierarchical level (Dubchak et al., 1999). These levels include family, super family, fold and class. *Family* (clear evolutionary relationship) consists of those proteins that

- a the sequence similarity among proteins is significant
- b proteins with almost same function and structures.

Super family (probable common evolutionary origin) consists of families where proteins have low sequence similarities and the existence of possible common evolutionary origin based on their structures and functional characteristics. *Fold or common fold* (major structural similarity) contains families and super-families with proteins having

- a same major secondary structures in same arrangement
- b same topological interconnection
- c less evolutionary relatedness.

Class gathers different folds that represent the protein secondary structure.

The major categories of *classes* are α (proteins with α -helices), β (proteins with β -sheets), α/β (protein structures formed by α -helices and β -sheets) and $\alpha + \beta$ (protein structures with large segregation of α -helices and β -sheets). Other categories are multi-domain, small proteins and peptides and few more.

Proteins sequences having sequence similarity between any two proteins with 35% at maximum and that has at least

80 residues were selected. According to the SCOP classification, each protein is associated with a fold. SCOP database has number of different folds but the major ones are (α , β , α/β , and $\alpha + \beta$). A *class* consists of a group of folds. SCOP contains 128 folds with one or more proteins, whereas after removing single-protein folds, the final dataset include 27 folds with four classes.

Table 1 describes the SCOP dataset with 27 folds (1..27), fold index, fold name, number of proteins in a fold, total number of folds and total number of proteins in a particular class and a class name. There are 698 proteins of the SCOP dataset.

The feature vectors of the SCOP dataset is extracted by transforming amino acids sequence into a sequence of six structural or physico-chemical properties of residues (Dubchak et al., 1997). The properties are amino acids composition (C), predicted secondary structure (S), hydrophobicity (H), normalised van der Waals volume (V), polarity (P) and polarisability (Z).

Twenty amino acids are divided into three different groups as polar (p), neutral (n) and hydrophobic (h). Table 2 provides details about the six properties/features of proteins such as C, S, H, V, P, and Z, symbol, dimension and the twenty amino acids placed in their corresponding group.

The dimension of each protein feature is 21 except the amino acid composition (C) which has the dimension 20 showing the percent existence of each twenty amino acid in a protein sequence. The remaining five features are based on the three descriptors composition (C) describing the percent composition of amino acids in protein sequence belonging to each one of the three groups; transition (T) illustrating the transition frequencies from p-n, n-h, h-p and vice versa; and distribution (D) showing the distribution pattern of amino acid at five different locations for each group with tentative starting position 0, plus 25% for each of the next location.

2.2 Ant colony optimisation

Dorigo (1992) initially introduced ACO algorithm motivated through the natural behaviour of ants. Aim of the algorithm was to find a shortest path based upon the ants behaviour searching for a path between a source of food and their residence (colony). With the passage of time, a number of extensions to the original concept have been made and implemented successfully for solving different types of problems. ACO is an iterative process in which selected population creates multiple possible solutions. ACO exploits the foraging behaviour of ant species working in such a way to search their food in less amount of time. The ants pour pheromone on the route to be followed by other ants of the colony (Dorigo et al., 2006).

Table 1 The SCOP dataset with 27 folds, fold index, fold name, number of proteins in a fold, and a class name

<i>Fold no.</i>	<i>Fold index</i>	<i>Fold name</i>	<i>No. of proteins</i>	<i>Class name</i>	<i>Total no. of proteins</i>	<i>No. of folds</i>
1	1	Globin-like	19	α	116	6
2	3	Cytochrome c	16			
3	4	DNA-binding 3-helical bundle	32			
4	7	Four-helical up-and-down bundle	15			
5	9	4-helical cytokines	18			
6	11	EF-hand	16			
7	20	Immunoglobulin-like beta-sandwich	74			
8	23	Cupredoxins	21			
9	26	Viral coat and capsid proteins	29			
10	30	ConA-like lectins/glucanases	13			
11	31	SH3-like barrel	16			
12	32	OB-fold	32			
13	33	Beta-trefoil	12			
14	35	Trypsin-like serine proteases	13	α/β	260	9
15	39	Lipocalins	16			
16	46	Beta/alpha (TIM)-barrel	77			
17	47	FAD (also NAD)-binding motif	23			
18	48	Flavodoxin-like	24			
19	51	NAD(P)-binding Rossmann-fold domains	40			
20	54	P-loop containing nucleotide triphosphate hydrolases	22			
21	57	Thioredoxin-like	17			
22	59	Ribonuclease H-like motif	24			
23	62	Alpha/beta-hydrolases	18			
24	69	Periplasmic binding protein-like	15			
25	72	Beta-grasp	15	$\alpha + \beta$	96	3
26	87	Ferredoxin-like	40			
27	110	Small; small inhibitors, toxins, lectins	41			

Table 2 The six protein properties, symbol, their dimensions and 20 amino acids in three groups

<i>Property</i>	<i>Symbol</i>	<i>Dimension</i>	<i>Polar (p)</i>	<i>Neutral (n)</i>	<i>Hydrophobic (h)</i>
Amino acids composition	C	20	-	-	-
Predicted secondary structure	S	21	α -helix	β -sheets	Turn
Hydrophobicity	H	21	R, K, E, D, Q, N	G, A, S, T, P, H, Y	C, V, L, I, M, F, W
Normalised van der Waals volume	V	21	G, A, S, C, T, P, D	N, V, E, Q, I, L	M, H, K, F, R, Y, W
Polarity	P	21	L, I, F, W, C, M, V, Y	P, A, T, G, S	H, Q, R, K, N, E, D
Polarisability	Z	21	G, A, S, D, T	C, P, N, V, E, Q, I, L	K, M, H, F, R, Y, W

ACO has proved its importance to the world because of its successful role in diverse applications since its inception, for example optimising the power network management problem (Abdelaziz et al., 2012), algorithms and problem complexity analysis (Ahangarikiasari et al., 2013), a mathematical problem (Anstreicher et al., 2002), and/or biomedical problems (Fidanova, 2006).

Real ants get attraction of human beings since long due to its wonderful social behaviour. The most promising feature is their queue formation between their source and destination while searching for the foods. The lesson learnt by most of the scientists, researchers from this behaviour of ants is to find the shortest paths in order to achieve their ultimate goal.

Biologists proved it experimentally that the way ants follow their antecedents is due to the communication between them through *pheromones* – chemicals with the ability to attract other individuals. The amount of pheromone concentration guides other ants to walk through the same path as pheromones disperse with the passage of time if, further, no pheromone addition takes place. The suitable path for the ants' family is the shorter one with more concentration as compared to the longer one. Researchers applied it for optimisation problems after thoroughly studying this behaviour. Solving variety of problems becomes the main reason of popularity of the ACO algorithms. Several ACO algorithms have been developed but the original one was called the ant system and is referenced by the algorithms suggested latter on Dorigo and Stützle (2004).

Since a decade, the scientists and researchers have greatly converted their attention towards ACO due to its high-rated success in discrete optimisation problems where other approaches have limitations. ACO always remains a best choice for these kinds of problems and finds the quality solution in a faster manner. Find a dynamic shortest path in a telecommunication network, job scheduling, image processing, bioinformatics and data mining are sample application areas to name.

2.3 Association rules discovery

A voluminous amount of data is gathered by many business enterprises on daily basis. Typical example of such data is market-based transaction, where many stores collect the customer information and the items list purchased at the checkout counters. Retailers analyse the data that help in their proper inventory management, marketing promotions, pricing, product placements and customer relationship management. Representing these relationships, association rules were introduced (Agrawal et al., 1993) in order to discover interesting relationship hidden in large databases.

An association rule is written as IF – THEN statement, where IF part of the rule is known as the antecedent (left-hand-side or LHS) and THEN part is said to be the consequent (right-hand-side or RHS) of the rule. Formally, association rule is the implication expression described as $X \rightarrow Y$, where the intersection of X and Y is NULL, and X and Y represent different sets of items. To observe the strength of the rule two factors are associated – support and confidence.

Support (s) is the fraction of transactions in a database that contain both X and Y , and confidence (c) is the occurrences of itemset in X that appear in transactions. To explain the concepts, suppose an example database with six items (I) and five transactions (T).

Itemset (I) = {bread, milk, eggs, butter, coke, fruit}

T1 = {bread, milk}

T2 = {bread, eggs, butter, coke}

T3 = {milk, eggs, butter, fruit}

T4 = {bread, milk, eggs, butter}

T5 = {bread, milk, eggs, fruit}

An example rule {milk, eggs} \rightarrow {butter} meaning that, if milk and eggs are bought, customers also buy butter.

To select the important rules from set of all possible rules, the two measures reflecting the significance and interest of the rules, that is, support and confidence is determined. Consider the rule {milk, eggs} \rightarrow {butter} with $X = \{\text{milk, eggs}\}$ and $Y = \{\text{butter}\}$.

- *support* (s) is represented as:

$$s = \sigma(X \cup Y) / N \quad (1)$$

- *confidence* (c) is represented as:

$$c = \sigma(X \cup Y) / \sigma(X). \quad (2)$$

Symbol σ is the support count (σ) showing the frequency of occurrence of itemset (I). For example;

$$\sigma(\{\text{milk, eggs, butter}\}) = 2.$$

$$s = \sigma(\text{milk, eggs, butter}) / N = 2/5 = 0.4$$

means that it occurs 40% of all transactions (two out of five transactions).

Confidence (c) of the rule {milk, eggs} \rightarrow {butter}

$$s = \sigma(\text{milk, eggs, butter}) / \sigma(\text{milk, eggs}) = 2/5 = 0.67$$

It means that 67% of the transactions containing milk and eggs the rule is correct. That is 67% of the times milk and eggs are bought, customers also buy butter.

The reason for using support and confidence is that

- 1 a rule with low support reflects its insignificance and less interested at business perspective and hence eliminated
- 2 confidence measure the reliability of a rule.

The goal of association rule mining is to discover all those rules that have support (s) \geq minsupp threshold and confidence (c) \geq minconf threshold, where minsupp and minconf are the corresponding support and confidence thresholds respectively.

Besides market-based transaction, association rule mining is also applicable in many areas of interest such as bioinformatics, medical diagnosis, web mining, text mining and scientific data investigation. One interesting type of association rule mining is the associative classification that discovers association rules based on the class labels where consequent part of the rule is always a class label. The dataset used for the discovery of association rules, include set of transactions each consist of a set of attributes and a predefined class label. Class association rule is represented as $X \rightarrow C$, where X contains set of attributes of a transaction and C is the class label. Along with support and confidence,

coverage is also specified in associative classification. It determines how much percent of the data is covered correctly by the rules.

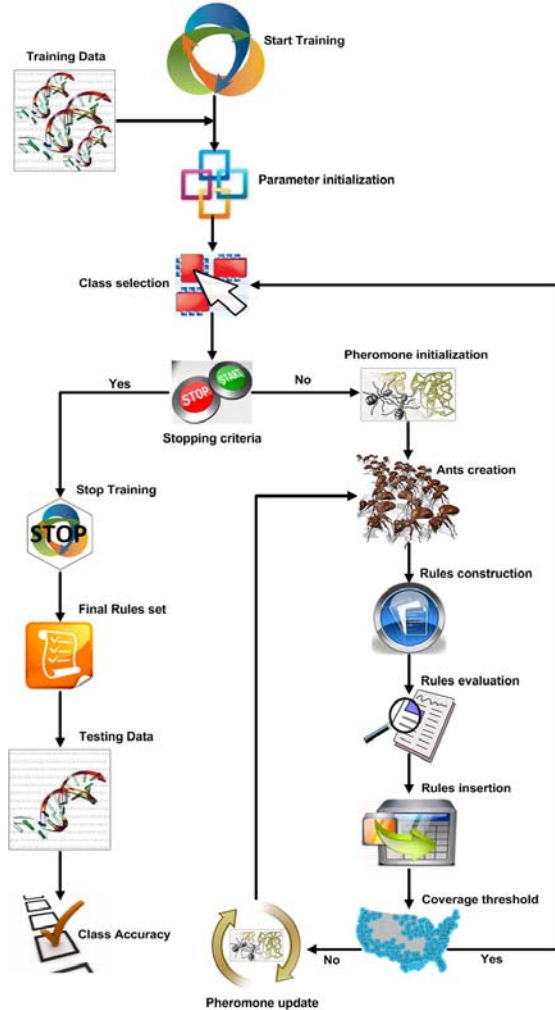
2.4 Working of the ACO-AC algorithm

Briefly describe the ACO-AC algorithm, it searches for the set of association rules using the dataset and build a classifier. It differs from conventional association rules mining that considers all possible rules and becomes unaffordable in large databases; instead, ACO-AC considers only the subset of such rules. Support and confidence are the two main components for the rule selection.

Figure 1 shows the working of the algorithm in detail. Rules searching are based on ACO mechanism. A graph represents a search space where nodes are the possible solutions.

A separate rules set is reserved and updated for a class during its rules generation process till the predefined minimum coverage threshold. The same process is repeated in the same style for rest for the class labels and completed when all classes have adequate rules.

Figure 1 Working of the ACO-AC algorithm (see online version for colours)



Training data and testing data in the figure are actually the protein sequences represented by feature vectors. Both are separate datasets contained in SCOP protein data. It is important to note that the feature vectors used for training are not used as testing data during experimentations.

2.4.1 Systematic sketch of the algorithm

- The moment training starts, the training data is provided followed by mandatory parameter initialisation such as rules set discovered (empty initially), minimum support, confidence, coverage, and maximum number of ants.
- A class is chosen from the set of classes for which association rules are generated and are placed after each generation in the rules list (empty initially).
- Pheromone values and heuristic function are initialised. The initialisation of pheromone values is made with the same amount at the beginning of rules construction for each class. The initial pheromone value (P) for the edge (i, j) is given as:

$$P_{ij}(t=1) = 1 / \sum_{i=1}^N v_i \quad (3)$$

The ij represent the edge with pheromone value P , N is the total number of attributes in a sample of a dataset except the class label, v_i represents the number of possible values at attribute N_i . Pheromone values that do not meet the minimum support are set to zero.

- Heuristic function measures the path quality between two items. Quality means the preference of the path to be selected by the ants. It is introduced in order to guide the ant to select the most attractive path and avoids probing the unnecessary search space. A heuristic function (h) used in this study is based on calculating the correlation of the next possible items and is of the form:

$$h_{ij} = \left(|x_i, x_j, C_k| / |x_i, C_k| \right) \cdot \left(|x_j, C_k| / |x_j| \right) \quad (4)$$

The term $|x_i, x_j, C_k|$ represent the number of uncovered training samples with items x_i, x_j and class C_k , divided by $|x_i, C_k|$ representing the uncovered samples having item x_i with class C_k in order to calculate the correlation between items x_i and x_j ; whereas, x_i and x_j are the items selected to be added in the rule. The next term $|x_j, C_k|$ shows the number of uncovered samples having item x_j with class C_k divided by $|x_j|$, which is the number of uncovered samples with item x_j . This part of the heuristic function shows the significance role of item x_j in determining the class C_k .

- During rule construction process, a counter (g) is used (initially 1) to control the addition of items up to its maximum number by an ant in the antecedent part of a rule. In the first generation, an ant construct a rule with single item, two-items rule in a second generation,

three-items rule in third and so on till N-items rule is constructed in the N^{th} generation. The addition of item (attribute-value pair) in the antecedent part of a rule is carried out on incremental basis by an ant and is selected primarily while ignoring all other attribute values of a rule. The selection probability (S) is shown as below.

$$S_{ij} = \left(P_{ij}(g) h_{ij}(c) \right) / \left(\sum_{i=1}^N A_i \sum_{j=1}^{v_i} [P_{ij}(g) h_{ij}(c)] \right) \quad (5)$$

where $P_{ij}(g)$ and $h_{ij}(c)$ represent the pheromone value and the heuristic function value between items (i, j) for the current generation and class respectively. The A_i is a value either '1' or '0' showing whether a particular attribute is accessed or not by the current ant. Items with high pheromone values have maximum chances for selection.

- Ants start constructing rules for each class based on the specified support and confidence criterion. The process continues until all the attributes are accessed or when minimum coverage criteria fulfils. After construction of rules by all the ants during a generation, they are evaluated and their quality is thoroughly examined. Rule quality (Q) is evaluated using the following criteria.

$$Q = TP / Coverage \quad (6)$$

where TP represent the number of training samples whose antecedent and consequent part is similar with the antecedent and consequent part of the ant's rule respectively. $Coverage$ shows the number of training samples that match with the antecedent part of the rule constructed by an ant. It tells about the confidence level of the rule and high confidence value resulting in a more accurate rule.

- Subsequent to rules construction by all ants, rules are placed temporarily and are examined so that rules not meeting the minimum support and confidence level, are discarded. The remaining rules are moved in a rule set specified for the current class (R_c) and are inserted in the discovered rules set (R_d) provided it enhances the quality of the latter set (R_d). It takes place in such a way that a rule from (R_c) is taken and compared with the existing one (R_d) one after the other. In case if the existing rules set (R_d) have greater or equal confidence than that of the newly constructed rule of (R_c) the insertion will not take place, otherwise it is included in (R_d). The process carries on till all the rules of (R_c) are compared.
- The rules discovery process for a specific class continues till it reaches the minimum coverage threshold. The same procedure is repeated for the next class and so on.
- In case the minimum coverage threshold is not yet fulfilled, the pheromone value is updated then. The

pheromone value for the edge (i, j) is updated on the paths through which ants passed after every generation.

$$P_{ij}(g+1) = (1-\epsilon)P_{ij}(g) + (1-(1/Q))P_{ij}(g) \quad (7)$$

where $P_{ij}(g)$ represent the pheromone value between items (i, j) for the current generation (g), ϵ is the pheromone evaporation coefficient and Q is the quality of the ant's rule constructed. More pheromone is deposited on the edge through which more ants' passes causing more probability for selecting an attractive edge.

- The training process finally stops after the completion of rules construction for all the classes.
- A large number of rules are inserted in the discovered rules set (R_d). There is a possibility of the existence of redundant rules in (R_d). These are the rules that are true for multiple training samples. Such rules are removed from (R_d) the final rules set.
- On the basis of confidence the final discovered rules set is sorted (highest to lowest) and is then used for classifying the test samples. Rules are tested one after the other until a rule becomes true for a testing sample. In this case the remaining rules are not tested. There also involve a pruning mechanism that notifies those rules, which are never used during the training. This way such rules are pruned from the final rules set and hence resulting a final classifier that is used for classifying the unseen samples. In other words, pruning increase the clarity of a classifier and enable them to classify a testing sample more accurately even with small number rules and in a faster manner.
- Finally, testing samples are presented to the classifier. The sample is assigned a class label of the rule with greater confidence and coverage range. Rules are tested for the test sample one after the other in a sorted manner. The rules, during this process, whose antecedent becomes true for a test sample, the consequent of the rule is responsible for assigning the resulting class. In case none of the rule happens to be true for a test sample, the default class is assigned, the majority class of the training set, as the final predicted class for the test sample.

3 Results and discussion

The ACO-AC algorithm is applied to the SCOP dataset used by (Dubchak et al., 1997). Experimentation has been carried out using the system with 2.6 MHz dual core processor and a memory of 2 GB. An analysis of the performance of the ACO-AC method and other various classification methods available in the Weka 3.6 software has been carried out in this study. Weka is a collection of machine learning algorithms for data mining tasks (Hall et al., 2009). The software contains multipurpose tools including classification. These methods include REPTree,

RandomForest, J48, FT, PART, JRip, MultiClassClassifier, LogitBoost, EnsembleSelection, bagging, IBK, SMO and NaiveBayes.

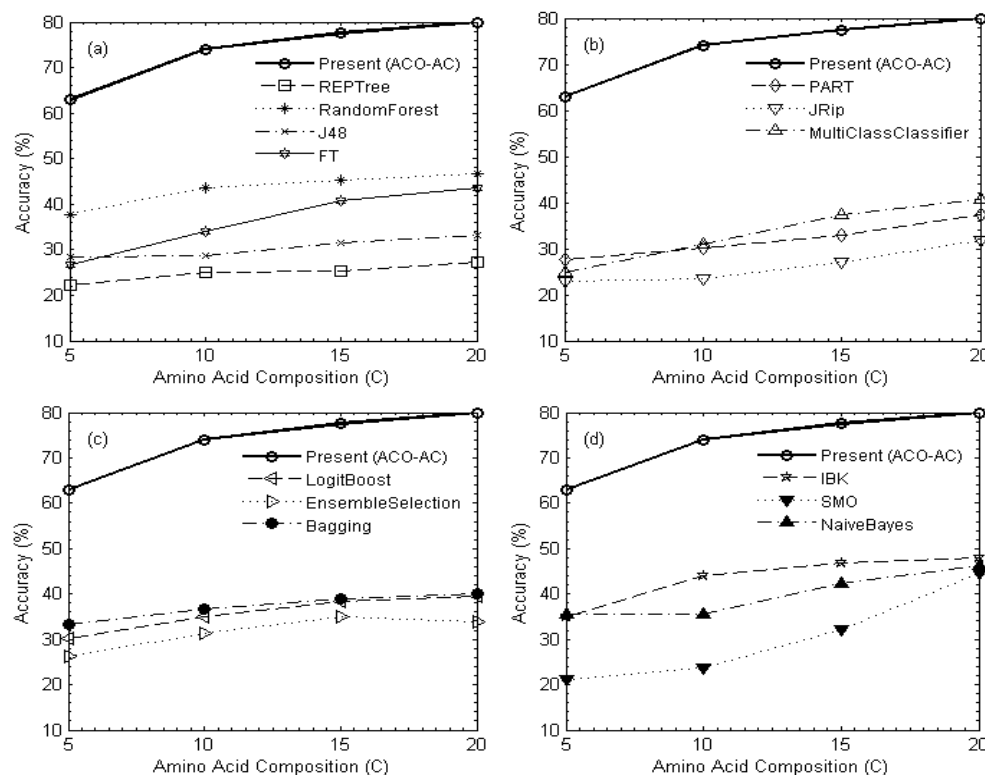
The comparison of protein fold classification accuracy is based on 10-fold discussed systematically using ACO-AC method and various methods available in the Weka 3.6 software. The distribution of proteins of all the classes in each fold is equal and not random. Results are generated using the best top 5, 10, 15 and all attributes of the six protein features such as C, H, S, P, V and Z using feature selection mechanism. The ACO-AC algorithm [referred as present (ACO-AC) in the figures] predicts better accuracy as compared to other available methods in terms of accurately classifying the protein fold using amino acid composition feature of protein.

3.1 Performance with amino acid composition (C)

In Figure 2(a), with top five best attributes, the classification accuracy of ACO-AC is 63% where as the maximum

accuracy against is of RandomForest method which is 38%. With ten best attributes, the ACO-AC method has an accuracy of 74% and outperforms the best among others – the RandomForest method, having an accuracy of 44%. With 15 best attributes, the present method has a high accuracy of 78% as compared to other available methods against which RandomForest appears with better accuracy of 45% among others available methods. Taking into account all the 20 attributes of amino acid composition feature of protein, the ACO-AC method comes up with an accuracy of 80% that is much better than all other available methods including the best among others, which is the RandomForest method with 47% accuracy. To summarise, with top best 5, 10, 15 and all (20) attributes of amino acid composition feature, the performance of the ACO-AC method is much better as compared to other available methods.

Figure 2 Classification accuracy (in percentage) for proteins feature – amino acid composition (C) of SCOP dataset, the performance of the ACO-AC algorithm against the other available classification methods, (a) open circle symbol with bold-solid-line represent the Present (ACO-AC) method which is identical in each plot; square box symbol with dashed-line represent REPTree method; six-crossed symbol with dotted-line represent RandomForest method; cross symbol with dash-dot-dash-line represent J48 method; and hexagon symbol with thick-solid-line represent FT method (b) diamond symbol with dashed-line represent PART method; downward-triangle symbol with dotted-line represent JRip method; and upward-triangle symbol with dash-dot-dash-line represent MultiClassClassifier method (c) left-triangle symbol with dashed-line represent LogitBoost method; right-triangle symbol with dotted-line represent EnsembleSelection method; and filled-circle symbol with dash-dot-dash-line represent bagging method (d) pentagon symbol with dashed-line represent IBK method; filled-downward-triangle symbol with dotted-line represent SMO method; and filled-upward-triangle symbol with dash-dot-dash-line represent NaiveBayes method



In Figure 2(b), with 5, 10, 15 and 20 attributes of amino acid composition feature of proteins, the ACO-AC method has an accuracy of 63%, 74%, 78% and 80% respectively. Against it, PART method has the maximum accuracy of 28% with 5 attributes; with 10, 15 and 20 attributes MultiClassClassifier method comes up with the maximum accuracy of 31%, 37% and 41% respectively. In brief, the ACO-AC method clearly shows better results compared to other available methods.

In Figure 2(c), maximum accuracy with 5, 10, 15 and 20 attributes among other available methods is the bagging method with 33%, 37%, 39% and 40% accuracy and is far behind than the present method with classification accuracy of 63%, 74%, 78% and 80% respectively.

In Figure 2(d), the classification accuracy of the ACO-AC method with 5, 10, 15 and 20 attributes is 63%, 74%, 78% and 80% respectively. NaiveBayes method has an accuracy of 36% with 5 attributes; IBK method with 10, 15 and 20 attributes emerges with better accuracy of 44%, 47% and 48%. Comparing the present versus the rest, the performance of the ACO-AC method is excellent.

3.2 Performance with hydrophobicity (H)

The performance of the ACO-AC method in terms of classification accuracy with hydrophobicity (H) feature of

protein is commendable in comparison with other available methods.

In Figure 3(a), ACO-AC method has a much higher accuracy of 61%, 73%, 75% and 77% against which RandomForest has the maximum accuracy of 36%, 36%, 36% and 35% with 5, 10, 15 and all (21) attributes respectively.

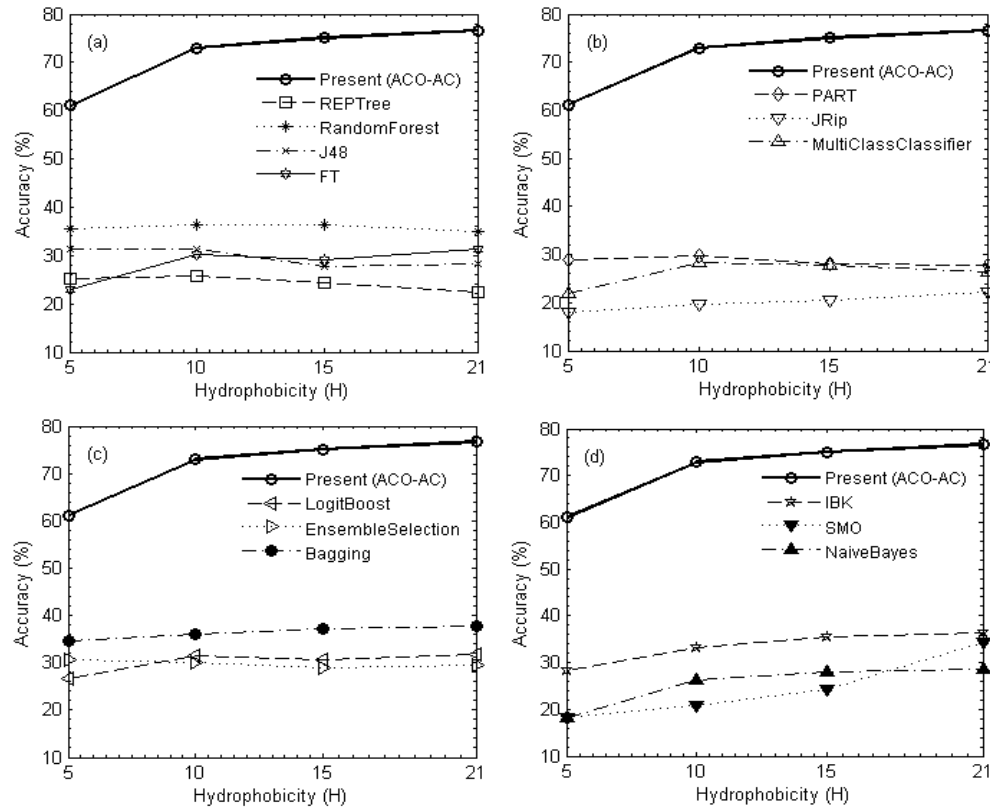
In Figure 3(b), with 5, 10, 15 and 21 attributes, against the accuracies of the present method, PART method comes up with the maximum accuracies of 29%, 30%, 28% and 28% respectively, which are far behind than the ACO-AC method.

In Figure 3(c), bagging emerges as the best one among other methods with 5, 10, 15 and 21 attributes having accuracies of 35%, 36%, 37% and 38% respectively against the present method. The present method outperforms all the other methods.

In Figure 3(d), as compared to the present method, the other methods have far less behind in performance. With 5, 10, 15 and 21 attributes, the best one method among other available methods against the present is the IBK method with 28%, 33%, 35% and 36% accuracies respectively.

In short, ACO-AC method performs much better with any number of attributes as compared to other available methods.

Figure 3 Classification accuracy (in percentage) for proteins feature – hydrophobicity (H) of SCOP dataset



Notes: The performance of the ACO-AC algorithm against other available classification methods. Same symbols description as in Figure 2.

3.3 Performance with polarity (P)

Another protein feature is the polarity (P) in which the ACO-AC method performed tremendously in terms of classification accuracy as long as the other available methods are concerned.

In Figure 4(a), the ACO-AC method with 5, 10, 15 and all (21) attributes of the polarity feature of protein produces the accuracies of 59%, 67%, 70% and 72% respectively. Against the same number of attributes, among other available methods the best one is the RandomForest with accuracies of 33%, 36%, 36% and 38% respectively. The present method outclasses all others methods.

In Figure 4(b), PART comes up with better performance having accuracies of 26%, 25%, 30% and 26% with 5, 10, 15 and 21 attributes respectively, which is very much lower than the present method.

In Figure 4(c), having 29%, 33%, 35% and 36% accuracies with 5, 10, 15 and 21 attributes, bagging among the other methods rises up with better result but lower than the present method.

In Figure 4(d), against the present method, IBK is the best one among others with accuracies of 26%, 31%, 34%, and 35% respectively. However, the ACO-AC is much better in classifying the protein folds.

3.4 Performance with predicted secondary structure (S)

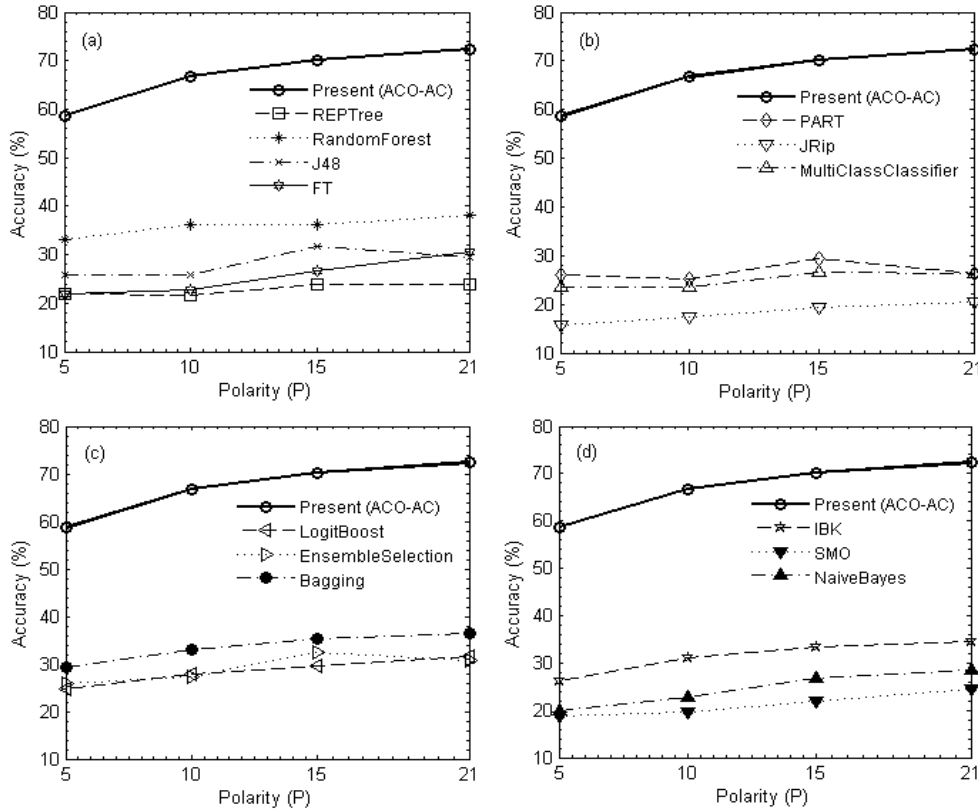
The ACO-AC method performance using the predicted secondary structure (S) feature of protein is extremely admirable because of producing the most encouraging results.

In Figure 5(a), the ACO-AC produces an accuracy of 80%, 84%, 82% and 82% with 5, 10, 15 and all (21) attributes of the predicted secondary structure feature. In contrast, the best one among other methods is the RandomForest method that produces 49%, 52%, 50% and 51% accuracy respectively. ACO-AC method outperforms all the methods with their best results.

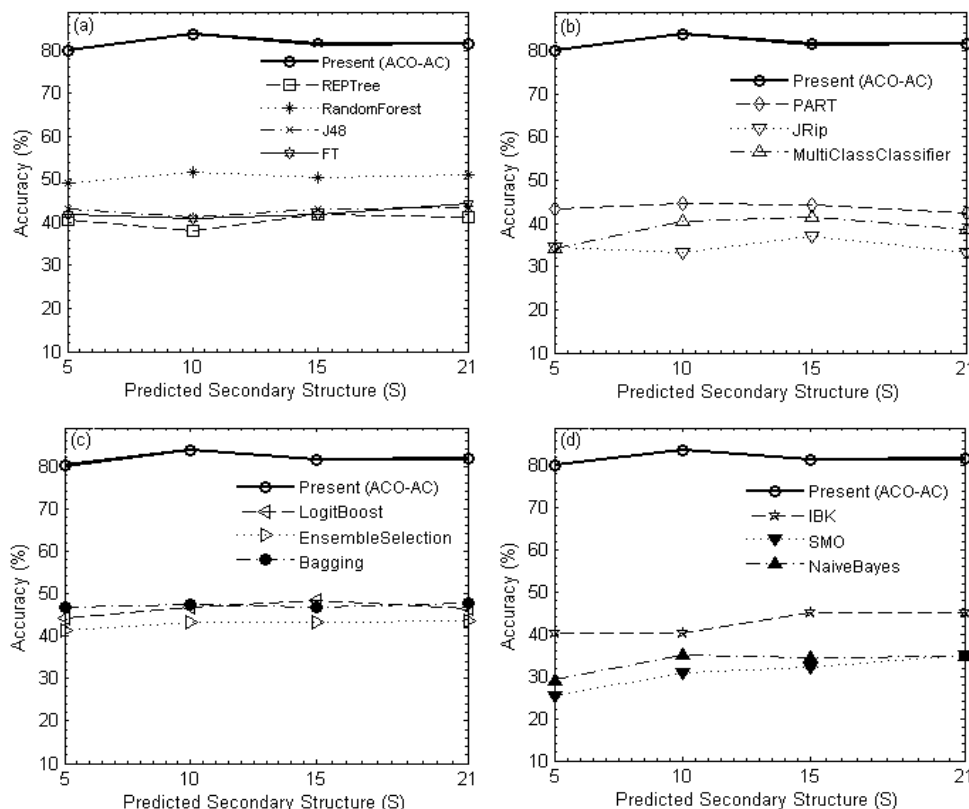
In Figure 5(b), against the accuracies of the ACO-AC method with 5, 10, 15 and 21 attributes, PART comes up with accuracies of 43%, 45%, 44% and 42% respectively, which is much lower than the present approach.

In Figure 5(c), bagging method, with 5, 10 and 21 attributes, emerges having the maximum accuracies of 47%, 47% and 48% respectively and LogitBoost produces 48% accuracy with 15 attributes against the accuracy of ACO-AC method. Both these methods have negligible performance against the present approach.

Figure 4 Classification accuracy (in percentage) for proteins feature – polarity (P) of SCOP dataset.



Notes: The performance of the ACO-AC algorithm against other available classification methods. Same symbols description as in Figure 2.

Figure 5 Classification accuracy (in percentage) for proteins feature – predicted secondary structure (S) of SCOP dataset

Notes: The performance of the ACO-AC algorithm against other available classification methods. Same symbols description as in Figure 2.

In Figure 5(d), with 5, 10 and 21 attributes, the present approach has much higher accuracies than the best among others, which is the IBK with accuracies of 40%, 40%, 45% and 45% respectively. In other words, with predicted secondary structure feature of protein ACO-AC performance is very much outstanding as compared to other methods.

3.5 Performance with normalised van der Waals volume (V)

Using the normalised van der Waals volume (V) feature also produces much better results than the competing techniques.

In Figure 6(a), the ACO-AC method has the classification accuracies of 54%, 68%, 72% and 75% with 5, 10, 15 and all (21) attributes of respectively. In contrast RandomForest method comes up with better accuracies though less significant than the present method with 32%, 35%, 35% and 38% respectively among other methods.

In Figure 6(b), against the ACO-AC method, with 5, 15 and 21 attributes, MultiClassClassifier method comes up

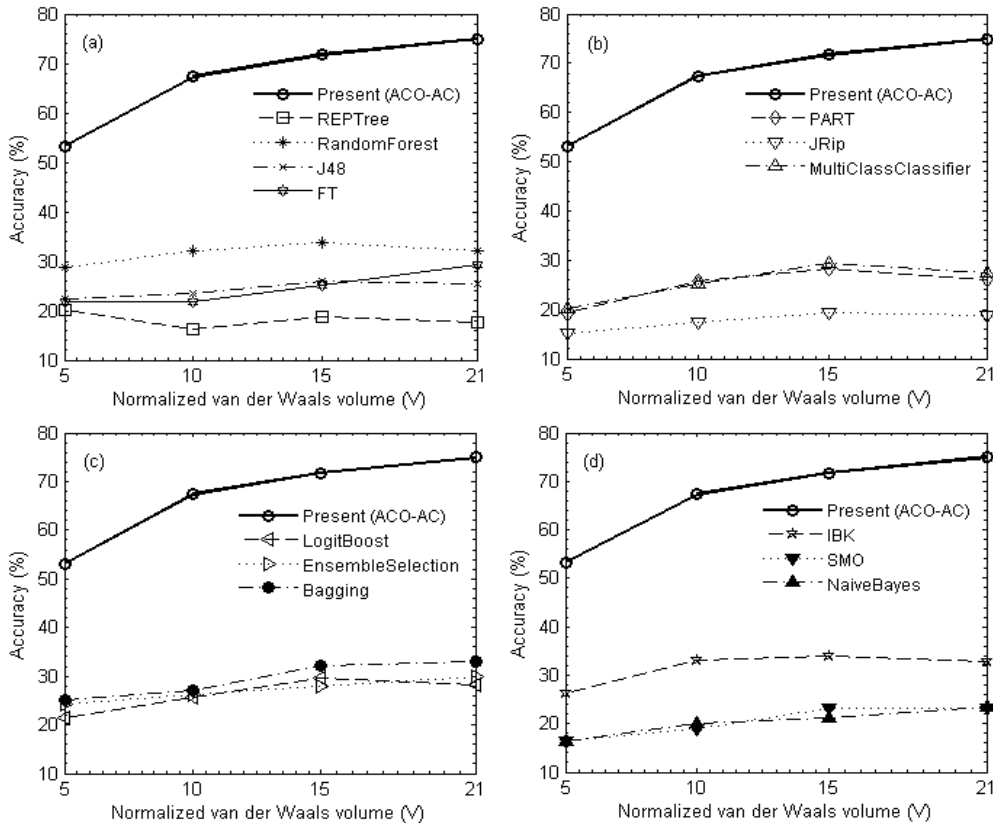
with their best results of 20%, 30% and 27% respectively; and with 10 attributes PART method has the better accuracy of 26% among others methods. In all cases, the present method outclasses all the other methods.

In Figure 6(c), bagging method has the better accuracies of 25%, 27%, 32% and 33% with 5, 10, 15 and all (21) attributes respectively. However, against the ACO-AC method it has the poorer performance.

In Figure 6(d), with 5, 10, 15 and 21 attributes, the best among other methods is the IBK with accuracies of 26%, 33%, 34% and 33% respectively. As compared to the present method, IBK is far behind. Simply, the ACO-AC method performance in terms of classification is much more than the other available methods.

3.6 Performance with polarisability (Z)

The performance of all other methods against the ACO-AC method using the polarisability (Z) feature of protein is almost negligible.

Figure 6 Classification accuracy (in percentage) for proteins feature – normalised van der Waals volume (V) of SCOP dataset

Notes: The performance of the ACO-AC algorithm against other available classification methods. Same symbols description as in Figure 2.

In Figure 7(a), the ACO-AC method produces 61%, 69%, 73% and 74% classification accuracies with 5, 10, 15 and all (21) attributes of polarisability (Z) feature respectively. RandomForest method comes up with better accuracies among other methods with accuracies of 32%, 35%, 35% and 38% respectively, which is less significant than the present method.

In Figure 7(b), PART method produces the better results with 5, 10 and 21 attributes having 27%, 26% and 27% accuracies and MultiClassClassifier method with 15 attributes produces the better accuracy of 29%. Both these methods are much lower in performance as compared to the ACO-AC method.

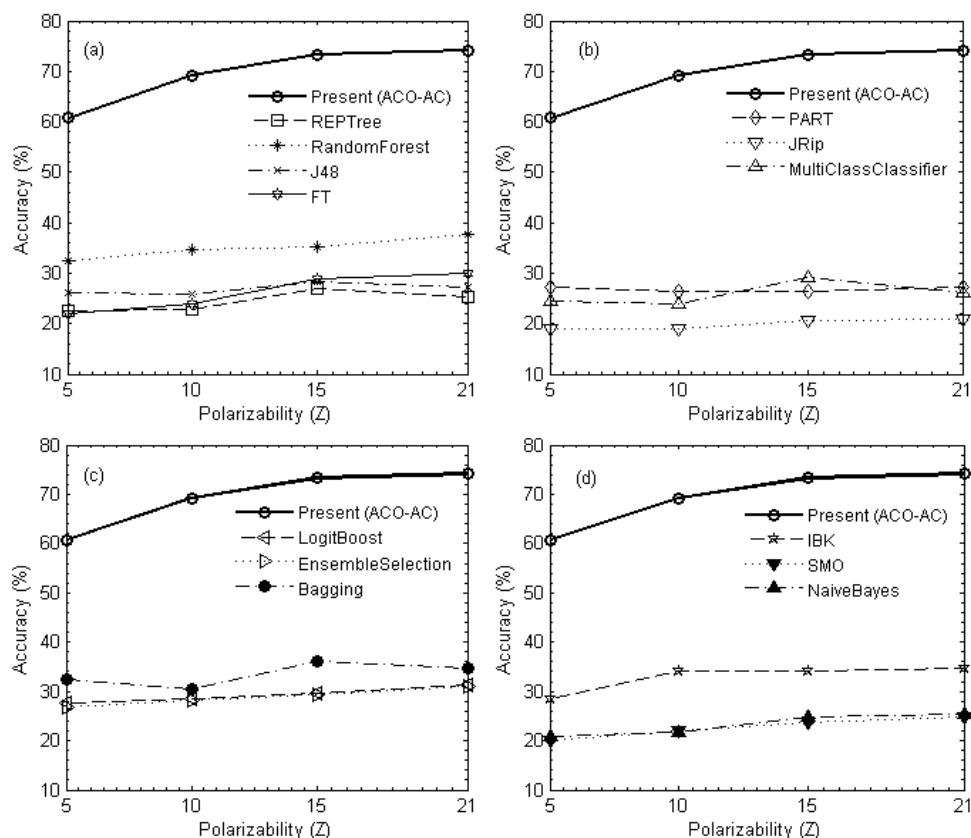
In Figure 7(c), against the performance of the present method with 5, 10 and 21 attributes, bagging method emerges with their best accuracies of 32%, 31%, 36% and 35% respectively. As compared to the ACO-AC method, all other methods have the poorer performance including the bagging method.

In Figure 7(d), with 5, 10 and 21 attributes, the best method among other methods against ACO-AC is the IBK method with accuracies of 28%, 34%, 34% and 35% respectively. However, all other methods including IBK are far behind in terms of classification accuracy. In summary,

ACO-AC method outclasses all other methods in all the cases.

4 Conclusions

The two important data mining techniques – association rules mining and classification rule discovery, used to expose the relations among large sets of data items. Rules discovery is supervised in nature. They have been adopted successfully in biomedical research for its significant outcome. In this study, the problem of classification of proteins into various folds is investigated and applied association rule mining techniques; the ACO-AC algorithm. The algorithm searches only for a subset of significant association rules rather than searching for all possible rules. It has the capability to deal with complex search space efficiently. Rules are discovered and evaluated at each generation, as a result quality rules are discovered in the subsequent generations. The challenge of computational complexity for large databases in this algorithm is efficiently handled because of mining a small set of association rules.

Figure 7 Classification accuracy (in percentage) for proteins feature – polarisability (Z) of SCOP dataset

Notes: The performance of the ACO-AC algorithm against other available classification methods. Same symbols description as in Figure 2.

The results produced by the ACO-AC using SCOP proteins dataset are compared with thirteen various classification methods available in the Weka 3.6 software. The dataset consists of 27 folds with 698 proteins. Results for all methods included in this study for protein fold classification accuracy is based on 10-fold. All the features of proteins data are categorised into four groups of best 5, 10, 15 and all attributes using feature selection mechanism. Experimental results show remarkable performance of ACO-AC as compared to other state of the art classification methods. Significant improvement has been noted in classification accuracy by the ACO-AC algorithm and outperforms other methods. In the Amino acid composition feature of protein the best-chosen method is the RandomForest method with five attributes, and IBK with 10, 15 and 20 attributes but are very much lower in classification accuracies than ACO-AC. Likewise, with the hydrophobicity feature of protein RandomForest method appears as better with five and ten attributes and bagging method with 15 and 21 attributes but outclassed by ACO-AC. Also, in polarity and predicted secondary structure feature of protein, the ACO-AC performance is outstanding as compared RandomForest method, better among other methods. ACO-AC in normalised van der Waals volume has far better results than both RandomForest with 5 attributes and IBK with 10, 15 and 21 attributes.

Similarly, in the polarisability feature of protein become negligible when compared with RandomForest method appears better in accuracies with 5, 10 and 21 attributes and bagging method having better accuracy with 15 attributes, both are far behind than the performance of the ACO-AC method.

To conclude the discussion, experimental results further reveal that the other available methods classify the correct fold of a protein almost by half, which is intolerable as long as such an important area of research is concerned. Second, as the number of attributes increases the performance of the ACO-AC method increases and the trend remains the same in all the six protein features except with 15 attributes of the predicted secondary structure feature. The behaviour divert again later on. Moreover, with predicted secondary structure feature of protein the accuracy rises up amazingly, which indicate the significance of the secondary structures in the formation of a protein.

To wrap up, the ACO-AC algorithm is more efficient and effective in the sense that it builds a robust learning model by discovering rules that are more significant increasing the coverage capability of the rules set and enhancing the classification accuracy convincingly even in complex search spaces. In future, the accuracy of protein fold classification can be increased by combining various proteins features such as CH, CHS, CHSP, and CHSPVZ

and this fact is already exposed in previous studies. It is also aimed that the algorithm shall deal all types of attributes instead of only categorical. Efforts will also be made to automate the minimum support and confidence threshold for the association rules discovery.

References

- Abdelaziz, A.Y., Osama, R.A. and Elkhodary, S.M. (2012) 'Application of ant colony optimization and harmony search algorithms to reconfiguration of radial distribution networks with distributed generations', *Journal of Bioinformatics and Intelligent Control*, Vol. 1, No. 1, pp.86–94.
- Agrawal, R., Imielinski, T. and Swami, A. (1993) 'Mining association rules between sets of items in large databases', *ACMSIGMOD Proc. ICMD*, pp.207–216.
- Ahangarikiasari, H., Saraji, M.R. and Torabi, M. (2013) 'Investigation of code complexity of an innovative algorithm based on ACO in weighted graph traversing and compare it to traditional ACO and Bellman-Ford', *Journal of Bioinformatics and Intelligent Control*, Vol. 2, No. 1, pp.73–78.
- Anstreicher, K.M., Brixius, N.W., Goux, J-P. and Linderroth, J. (2002) 'Solving large quadratic assignment problems on computational grids', *Mathematical Programming*, Vol. 91, No. 3, pp.563–588.
- Basu, S., Bhattacharyya, D. and Banerjee, R. (2013) 'SARAMA: a standalone suite of programs for the complementarity plot – a graphical structure validation tool for proteins', *Journal of Bioinformatics and Intelligent Control*, Vol. 2, No. 4, pp.321–323.
- Bitello, R. and Lopes, H.S. (2006) 'A differential evolution approach for protein folding', *CIBCB Proc.*, pp.1–5.
- Chen, P., Liu, C., Burge, L., Mahmood, M., Southerland, W. and Gloster, C. (2009) 'Protein fold classification with genetic algorithm and feature selection', *Journal of Bioinformatics and Computational Biology*, Vol. 7, No. 5, pp.773–788.
- Cheng, J. and Baldi, P. (2006) 'A machine learning information retrieval approach to protein fold recognition', *Bioinformatics*, Vol. 22, No. 12, pp.1456–1463.
- Cheng, J., Tegge, A.N. and Baldi, P. (2008) 'Machine learning methods for protein structure prediction', *IEEE Rev. Biomed. Engg.*, Vol. 1, No. 10, pp.41–49.
- Dill, K.A., Ozkan, S.B., Shell, M.S. and Weikl, T.R. (2008) 'The protein folding problem', *Ann. Rev. of Biophy.*, Vol. 37, No. 1, pp.289–316.
- Ding, C.H.Q. and Dubchak, I. (2001) 'Multi-class protein fold recognition using support vector machines and neural networks', *Bioinformatics*, Vol. 17, No. 4, pp.349–358.
- Doddi, S., Marathe, A., Ravi, S.S. and Torney, D.C. (2001) 'Discovery of association rules in medical data', *Med. Inform. Internet. Med.*, Vol. 26, No. 1, pp.25–33.
- Dorigo, M. (1992) *Optimization, Learning and Natural Algorithms*, PhD thesis, Politecnico di Milano, Italy.
- Dorigo, M. and Stützle, T. (2004) *Ant Colony Optimization*, MIT Press, MA, USA.
- Dorigo, M., Birattari, M. and Stützle, T. (2006) 'Ant colony optimization – artificial ants as a computational intelligence technique', *IEEE Comp. Int. Mag.*, Vol. 1, No. 4, pp.28–39.
- Du, Q., Wei, D. and Chou, K.C. (2003) 'Correlations of amino acids in proteins', *Peptides*, Vol. 24, No. 12, pp.1863–1869.
- Dubchak, I., Muchnik, I. and Kim, S-H. (1997) 'Protein folding class predictor for SCOP: approach based on global descriptors', *ICISMB Proc.*, pp.104–107, AAAI Press.
- Dubchak, I., Muchnik, I., Mayor, C., Dralyuk, I. and Kim, S-H. (1999) 'Recognition of a protein fold in the context of the SCOP classification', *Proteins: Structure, Function and Bioinformatics*, Vol. 35, No. 4, pp.401–407.
- Fairchild, S., Pachter, R. and Perrin, R. (1995) 'Protein structure analysis and prediction', *The Math. J.*, Vol. 5, No. 4, pp.64–69.
- Fidanova, S. (2006) '3D HP protein folding problem using ant algorithm', *BioPS Proc.*, pp.III.19–III.26.
- Giannopoulou, E.G. (2008) *Data Mining in Medical and Biological Research*, Vienna, Austria.
- Grosan, C., Abraham, A. and Chis, M. (2006) 'Swarm intelligence in data mining', in Abraham, A. et al. (Eds.): *Swarm Intelligence in Data Mining, Studies in Computational Intelligence*, Springer, Vol. 34, pp.1–20.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H. (2009) 'The WEKA data mining software: an update', *SIGKDD Explorations*, Vol. 11, No. 1.
- Hart, W.E. and Newman, A. (2005) 'Protein structure prediction with lattice models', in Aluru, S. et al. (Eds.): *Handbook of Computational Molecular Biology*, Chapman & Hall/CRC, USA.
- Hu, X-M., Zhang, J., Xiao, J. and Li, Y. (2008) 'Protein folding in hydrophobic-polar lattice model – a flexible ant-colony optimization approach', *Protein and Peptide Letters*, Vol. 15, No. 5, pp.469–477.
- Huang, C.D., Lin, C.T. and Pal, N.R. (2003) 'Hierarchical learning architecture with automatic feature selection for multiclass protein fold classification', *IEEE Tran. NanoBio*, Vol. 2, No. 4, pp.221–232.
- Hubbard, T.J.P., Murzin, A.G., Brenner, S.E. and Chothia, C. (1997) 'SCOP: a structural classification of proteins', *Nucleic Acids Research*, Vol. 25, No. 1, pp.236–239.
- Igel, C., Gebert, J. and Wiebringhaus, T. (2004) 'Protein fold class prediction using neural networks with tailored early-stopping', *IJCNN Proc.*, pp.1693–1697.
- Kin, T., Kato, T. and Tsuda, K. (2004) 'Protein classification via kernel matrix completion', in Schölkopf, B. et al. (Eds.): *Kernel Methods in Computational Biology*, MIT Press, MA, USA.
- Kotlyar, M. and Jurisica, I. (2006) 'Predicting protein-protein interactions by association mining', *Info. Sys. Front.* Vol. 8, No. 1, pp.37–47.
- Langlois, R.E., Diec, A., Dai, Y. and Lu, H. (2004) 'Kernel based approach for protein fold prediction from sequence', *IEEE ICEMBS*, Vol. 2, pp.2885–2888.
- Langlois, R.E., Diec, A., Perisic, O., Dai, Y. and Lu, H. (2006) 'Improved protein fold assignment using support vector machines', *Int. J. Bioinfo. Res. App.*, Vol. 1, No. 3, pp.319–334.
- Lin, K-L., Lin, C-Y., Huang, C-D., Chang, H-M., Yang, C-Y., Lin, C-T., Tang, C.Y. and Hsu, D.F. (2007) 'Feature selection and combination criteria for improving accuracy in protein structure prediction', *IEEE Tran. NanoBio*, Vol. 6, No. 2, pp.186–196.
- Liu, N. and Wang, T. (2007) 'A simple method for protein structural classification', *J. Mol. Graphics Modeling*, Vol. 25, No. 6, pp.852–855.

- Marsolo, K., Parthasarathy, S. and Ding, C. (2005) 'A multi-level approach to SCOP fold recognition', *IEEE Symp. Bioinfo. & Bioengg. Proc.*, pp.57–64.
- Okun, O. (2004) 'Feature normalization and selection for protein fold recognition', *Fin. AIC (STeP) Proc.*, pp.207–221.
- Shahzad, W. and Baig, A. (2011) 'Hybrid associative classification algorithm using ant colony optimization', *IJICIC*. Vol. 7, No. 12, pp.6815–6826.
- Shen, H-B., Chou, K-C. and Cr, K.A. (2006) 'Ensemble classifier for protein fold pattern recognition', *Bioinformatics*, Vol. 22, No. 12, pp.1717–1722.
- Sree, P. K., Babu, I. R. and Devi, N. U. (2013) 'PSMACA: an automated protein structure prediction using MACA (Multiple Attractor Cellular Automata)', *Journal of Bioinformatics and Intelligent Control*, Vol. 2, No. 3, pp.211–215.
- Taguchi, Y-h. and Gromiha, M.M. (2007) 'Protein fold recognition based upon the amino acid occurrence', *PRBI, LNCS*, Vol. 4774, pp.120–131.
- Yang, B., Sui, H., Wu, Q. and Wang, L. (2010) 'A data mining approach to predict protein secondary structure', in *ICCA SM 2010: Proceedings of the International Conference on Computer Application and System Modeling (ICCA SM)*, Taiyuan, China, pp.V6-589–V6-593.