Protein Function Prediction Model Using Multilabel Classification Algorithm

E.Balraj, S.Priyanka, B.Preethi

Abstract— ***The Ant Colony Optimization algorithm is helpful for predicting protein functions. It is the combination of novel concept about Optimization of Ant Colony algorithm and classification of hierarchical multi-label. This classification method is represented in hierarchical structures. In this methodology, each class has multiple class labels and these labels are defined in the form of Trees or Directed Acyclic Graph. This problem mainly focuses on the area of Bio-metrics. As there is a amount of increase which categories as uncharacterized proteins that can be available to perform prediction and analysis. It is necessary for resolving protein functions in order to which can be increase a trending biological intelligence. A protein can perform many functions and definition schemes. Those schemes are organized in structures based on the classifications of hierarchy. A common problem that arises when using a conventional flat classification approach is that it cannot predict for multiple class labels as it does only for a class label. This proposed Ant Colony Optimization(ACO) discovers the proper list for multiple class labels.***

Index Terms— ***Ant colony Optimization, Multi Label Classification, Machine learning, Protein prediction, Bio-metric Analysis.***

I.INTRODUCTION

Ant Colony Optimization (ACO) is a population-based approach by defines behavior of ants to solve combinatorial problems. ACO is an iterative process by using the pheromone which produces by ants that solution has been detected. This process is probabilistically guided by heuristic information which is given by ants to solve the problem by using the pheromone and storage of shared memory containing experience of pheromone gathered by previous iterations. A classification is a common approach to data mining actions and task. The main goal of the optimization which can analyze a relationship between input and output values. A set of data defined as the classified state, where every data set is explained to predictor attribute with a class attribute values. The Classification of approach can be categories into two stage of types. One is a Single Label Classification algorithm and another Multi-Label Classification. It consists of several two stages. The first stage which includes the label as a group of data and allow the process with a predefined label as user defined set of input, an exemplary should classifies and also defines the accord among predictor values and defined value of class attribute. The next stage is about classified model whereas classifies unknown dataset with unknown class values. each data in the set can be accomplice with value of one class value attribute and those class values which can be unrelated is known as single label classification. The single label classification problems are usually referred to as non-hierarchical. Multi-label classification is a classification action where an instance of more than one existing classes can be performed a classified operation. Label data extracted information from various domains, like web pages, text, multimedia (audio, image, videos) and biology. The problem can be the focus on classification about hierarchical multi-label. Each data set should accomplice with more than one values of class at coinciding and values of class can be organized as Trees or Directed Acyclic Graph which should be define as the hierarchical structure.

In multi-label classification, A collected data values should create into a different state of paradigm and every paradigm should be created with a whole of classified data. A task of Multi-Label classification is to predict the value sets of unknown paradigm by defines into performing realize to training data paradigm with defined state of labels. The hardest part about the classification of hierarchical by multi-label is to define the relationship among classes on nodes serves at a bottom level than classes serves on nodes at the top level. Where the defined known of data per class defines to be minimum stage at a lower level of hierarchy as disagree to a top-level stage of the hierarchy. The multiple unrelated classes problem can be boost about involvement of hierarchical classification, which can perform an operation in the undefined classification and protein level function prediction.

The subsequent of Scientists in the area of the research field, find and produce a huge set of increase uncharacterized proteins to generate analysis and create new biological information state by using Biometrics. It is significant to process hierarchical structure classification models which create ensured by defined values which can be describe attribute set also interact of values among protein features and related certain category. Concerning the issue of prediction of analysis protein, which classified stage where predicts the perform values correspond to various features and their classes correspond to various functions. The defined state of prediction of data values should achieve on multiple prediction of function values and their defined function which are constructed as a hierarchical function structures shows the example of defined schemes as FunCat and Gene Ontology which defines of classifies on this stage which is occurrence attribute issue about hierarchical multi-label.

Revised Manuscript Received on April 12,2019.

E.Balraj, Asst.Professor,Department of Information Technology, M Kumarasamy College of Engineering, Karur.

S.Priyanka, UG Student, Department of Information Technology,M Kumarasamy College of Engineering, Karur.

B.Preethi, UG Student, Department of Information Technology,M Kumarasamy College of Engineering, Karur.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

II. LITERATURE SURVEY

I.I. Ant Colony Optimization

An Ant colony optimization (ACO) is a population-based state that was used in different categories problems level defines difficult. Ant Colony Optimization algorithm is a very effective method to partitioning graphs can be found at the best solution. The algorithm can be simulated relationship between the behavior of artificial ants and real ants. To implement ACO, the optimization defines a creative operation of a problem which can find the best path on the graph. Each artificial ants which are participate in operation can build the new solution using the movement of each ant and every stage of pheromones can be stored in a storage. A Pheromone value is generated by ants can easily modify at runtime. An ACO algorithm can define the best solution as a result of the collaborative operation as optimization among certain different stages on the ant. An ant chooses of trail method also categories pheromone levels where an operation can be deposited by previous operation ants and also easily updated to finds the best solution on optimization.

By using example let us realize the Ant Colony Optimization. Let us assume, The application process as traveling salesman problem (TSP). In the Problem a set of defined data cities locations and distances among cities they are already defined as constraints. The problem which consists of a man which finds a nearest path of minimum length that man visits each and every city as only once. To implement ACO in TSP, let us visualize the graph which consists of data is the collection of nodes and vertices. Even though a problem is possible to move one vertex to another vertex, and the graph is entirely connected with the defined number of vertices and cities as equal. The categories value of classified Pheromone can be alter at runtime which can serves the patience of the ants by using colony optimization, where heuristic information values are the problem because those values are dependent in area of TSP, a data set can perform the inverted of the lengths.

I.II. MuLAM Optimization

Multi-Label Ant-Miner is known as recently discovers ACO algorithm, which names as MuLAM (Multi-Label Ant-Miner), for developed as classification of multi-label where defined set of rules. In basis, MuLAM various from already existed content of Ant-Miner classifies into three categories, as known as, First, a classification value can forecast multi class attributes values, where classification issues a model can exist into multi label class value. Second step on each repetition of MuLAM produces a data value of defined rules alternative of a rule value as in the exist Ant-Miner classification. On next stage, The pheromone defines in the form of matrix should contains class value of attribute and every update about pheromone occur on the form of every matrix where class values should serves the phase rule of consequent.

In a process to analysis with data on multi-label, MuLAM engage can create a class value whether multi class values should be predicted by the alike rule.

The stage of scheduled as brief on the various classification about Ant-Miner algorithm simple for some set of issues, called as hAnt-Miner (Hierarchical Classification Ant-Miner), handling rule structure of hierarchal analysis where data measure to predefined explanation on pheromone update and examining information switched for defined classification value , as well as an enhanced rule which serves to create hierarchically similarity classes of subsequent of a rule. A hAntMiner not able to hierarchical relate multi-label problems, where an every case belong to multiple classes which should not ancestor/descendant. The state focus on extends the idea of hAnt-Miner into the hierarchical multi-label classification issue class.

I.III. Gene Ontology

Genome annotation is processed can be obtaining data set from the genes and generate the product about gens where use the scheme as the GO ontology. The discussion belongs GO Ontology which contains gene annotation for combination and propagation on the Genome and their related website by using AmiGO. In extra detail of gene product and gene identifier is to introduce their relevant GO term, and their GO annotations by using the collected data. The remark can be realize the annotation and their evidence code representing which type of evidence analyze which the annotation is based and their function. The analyze of proof code introduces from defined vocabulary of codes covering both manual and automated annotation functions and their methods.

For Understand, TAS(Traceable Author Statement) refers a curator of scheme and action of read a statement which published their scientific paper and which includes the meta data. For Understand, ISS(Inferred from Sequence Similarity) refers a human curator of scheme which perform review the output from the analyze of a sequence related search and verifies that is very useful for biologically statement.

For Understand, IEA (Inferred from Electronic Annotation) contains the automatic process has introduced the code for annotations and their schemes. The recent survey of GO annotations were inferred as computationally over all 98% , not by curators. As the GO annotations are not checked by a human and their schemes of GO Consortium which contains annotations not much reliable and state adds the single subset in the data source available online in AmiGO.

Full annotation related data set can be download from the online Go website . Additionally, The algorithm of machine learning have been designed and implemented the process of prediction related to Gene Ontology annotations.

I.IV. Protein and its Structure

Proteins are the most essential molecule in our Human body and their cells and which plays important role in living organisms. By density, proteins are the group of a collective valuable component on density of human cells which can be affect in human cell and their functions. Every content of protein contains a separate function derived from cellular which supports to process on cellular locomotion and cell signaling. For total, there are various of seven categories of

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

proteins which includes antibodies, enzymes, hormones, some of the insulin and so. The protein contains as more diverse functions, which is common for typically derived from one data acid among 20 amino acids. The architecture of a protein may belong to globular functions and their sketch which helps for every protein with their specific function. We can make the parts as into various of architecture on protein molecules as two types of common parts: Those are fibrous proteins and globular proteins. Fibrous proteins are known as elongated and insoluble. There are of Globular proteins available for prediction For reference should define as compact, soluble, and spherical in their shape.

Globular and fibrous proteins should exhibit multiple types of their protein structure. There are main important categories of four levels of protein architecture: Those are primary, secondary, tertiary, and quaternary. The protein structure contains the variety of level in which one to another as the degree of a protein chain. A single layer content of protein molecule may include containing multiple protein structure and their category. The basic structure of a protein has explained its function. For Understand, A collagen has a super-coiled helical shape which is long, stringy, strong, and resembles a rope and provides great support. For Understand, Hemoglobin is a globular protein which can be describe as folded and compact. The spherical shape is useful for maneuvering by usage of blood vessels. Various Example, a protein also contains a non-peptide group. These are referred as cofactors and known as coenzymes which are from organic. A protein state is an inorganic group, known as a metal ion or iron-sulfur cluster.

III.EXISTING SYSTEM

hAnt-Miner algorithm is a new revelation to introduced hierarchical classification structure and data set rules in the design as IF antecedent THEN consequent. The rule of antecedent which can be contains as a conjunction of data values into the set where the statements based on predictor class attribute values. The rule of consequent which can contain as class labels in the form of the set of values in properly as various levels of the class attribute in form hierarchy structure which respecting into ancestor/descendant of defined attribute relationships. This algorithm can make into various parts of a rule which construct data into two types of different colonies. one ant colony is used to create the set of antecedent rules and various of ant colony is used to create the set of consequent rules, and both colonies can work as into cooperative fashion.

At the final stage, The construction rule process has completed, then rules phase can be created by the ants are pruned which to eliminate the unwanted terms where specifies asset of attribute conditions from their antecedent and state allows to create the observe terms in the regional search operator and class values from their referred consequent. The next stage, each level of pheromones is updated by using stored with every stage of iterations. The good rule which is based on measurement of quality in every stage of the current iteration compares to all iterations and where the pheromones are stored. The construction rule procedure is reworked until a existing set of iterations which is specified by a user to reach their iterations, or the state to perform as same which is already defined previous iterations is known as best so far rule. The rule best-so-far is introduced to the rule on their list and also covers the training data set. For Understand, the data set should delight their rules of conditions as antecedent and eliminate the unwanted from their training data set. In total, hAnt-Miner can be introduced as an algorithm of memetic which should be combined as conventional concepts and methods which includes the ACO metaheuristic information with data set should induction of various algorithms. An Introduced of their list with rules of classification with sequential covering state to covers the training data set.

Algorithm 1

```
input : training data set examples
output: discovered rule list
begin
  training set ←all training examples;
  rule list ←0/ ;
  while |training set| > max uncovered examples do
    rulebest ←0/ ;
    i←1;
    repeat
      rulecurrent ←0/ ;
      for j←1 to colony size do
        // use separate ant colonies for antecedent and consequent construction
        rulej ←CreateAntecedent()+CreateConsequent();
        // applies a local search operator
        Prune(rulej);
        // updates the reference to the best rule for each iteration
        if Q(rulej) > Q(rulecurrent) then
          rulecurrent ←rulej ;
        end
        j← j+1;
      end
      UpdatePheromones(rulecurrent );
      if Q(rulecurrent) > Q(rulebest) then
        rulebest ←rulecurrent ;
      end
      i←i+1;
    until i ≥ max number iterations OR RuleConvergence();
    rule list ←rule list +rulebest ;
    training set ←training set −Covered(rulebest ,training set);
  end
  eturn rule l
```

IV. PROBLEM DESCRIPTION

To perform the operation of analyze on data set in hAnt-Miner, We have some of the condition. The main problem issue on the hAntMiner is the information which is heuristic and also performs the operation of measurement of entropy and which is not suitable for classifies as hierarchical. The fact of a problem is identifying the relationship between the hierarchical classes. The hAnt-Miner is used to measure the entropy which is common for all class labels and perform

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

the calculation. Every class labels which perform the operation an individual should not consider the parent-child relationship among all class labels.

Another important drawback which is based on the rule of measurement the quality is to perform the overfitting. The evaluation is should be considered by the rule of small coverage based on the generic rules. Let us consider for understand, The class labels 1.2.1 with the set of data example as 20 and should have the proper rule set as the class label as 1.2.1 with the specified consequent. For rule1, The coverage of correctness with 5 set of examples out of a total 5 covered rule and rule2, which covers the correctness of data set as 19 out the total covered 20. Based on the situation, we have to analyze the high quality, because all the defined modules should be covered by classified correctness than rule2, which can be mismatch state classifies the data set for example, through rule2 which can cover all but the examples which it belongs to class 1.2.1 and should be focus area of the main issue that the measured quality on hAnt-miner could be modified as easily to remove the state of overfitting can be evaluated by the specific classes.

The approach of the drawback which favors the rules should be predicting class label at a top level of a hierarchy. The process can potentially save the already existing rules and also includes the specific set of the class label which is removed from training data set through the entire set process observed from already existing experiment result. In the last stage, The set should not allow processing multi-label data to construct the single path with the consequent rules. Let us focus on the part of protein prediction and which is able to perform the operation such as more than one. So, The process can arise the more important issue.

V. PROPOSED WORK & THEORETICAL RESULT

The hierarchical multi-label has introduced the classification by adopting ant colony algorithm which is referred as the Hierarchical Multi-Label Classification Ant-Miner(hmAnt-Miner). The Ant-Miner is establish to beaten some of the pre-defined limitations.

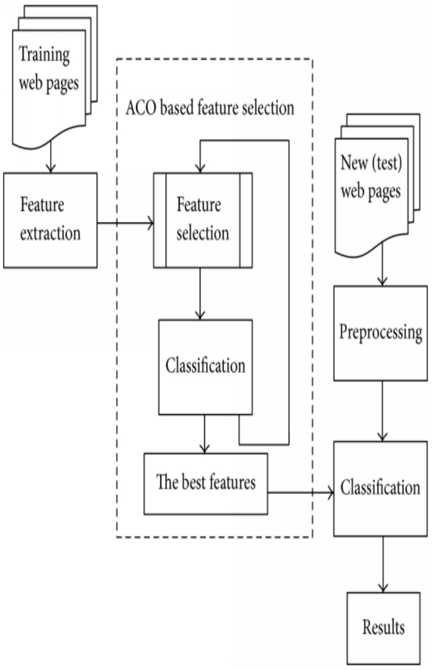
There is a huge difference between the hmAnt-Miner and hAnt-Miner algorithms based on the procedure.

The following procedure is the basic difference how the hAnt miner differs from the hmAnt-miner :

- ✓ The Rule of consequent can analyze the dataset of procedure which is based on the covered rule and allow as to create and predict the class label as more than once at process performed by the same time as multi-label rules. So, The result of construction graph represented that hmAnt-Miner uses the single construction graph on the creation only on the antecedent.
- ✓ The Distance measure the Euclidean formula can define information which is based on the function class of heuristic, where the dataset of each class represents by vector and membership values are defined in the Euclidean space. To perform the operation of measurement, we can use the distance instead of entropy in hAntMiner and help us to identify the relationship between the class labels. The Dataset of examples such as ancestor or descendant related to class labels will be more relative to the unrelated dataset as examples. To perform the operation on multi-label classification, the Approach is activated from the CLUS-HMC algorithm which is based on the decision tree induction compare to rule induction.
- ✓ To evaluate the rule quality we need to measure the distance and those are suitable evaluated measurement for hierarchical related multi-label problems.
- ✓ The procedure of rule pruning is not common for construction of consequent rule. It performs the operation of recalculation when the state of the antecedent is modified during pruning and already defined set of examples may be changed.

Algorithm:

```
        input : rule to be pruned
        output: the pruned rule
begin
        $rule_{best} \leftarrow rule_{current}$;
        $q_{best} \leftarrow Q(rule_{best})$;
        repeat
        $rule_i \leftarrow rule_{best}.antecedent - last\ term(rule_{best}.antecedent)$;
        calculate consequent($rule_i$);
        $q_i \leftarrow Q(rule_i)$;
        if ($q_i \geq q_{best}$) then
        $rule_{best} \leftarrow rule_i$;
        $q_{best} \leftarrow q_i$;
        end
        until $q_i < q_{best}$ OR $|rule_{best}.antecedent| = 1$ ;
        return $rule_{best}$;
end
```



Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

VI. CONCLUSION

The paper prompts about novel ant colony algorithm works based on the classification of hierarchical multi-label, which can be referred as Hierarchical Multi-Label Classification Ant-Miner(hmAnt- Miner). The already existing objective about classification algorithm of hierarchal which can discover a single label classification and forms can be ordered list as IF-THEN rules of classification which predict the whole hierarchy at least once. The data set of examples may be assigns to multiple unrelated classes. The information which is derived from the class hierarchy of hmAnt-Miner which user can measure the procedure of continuous attributes to construct the ACO graph. Hierarchical multi-label classification is the best method to perform the operation of measurement as entropy used in hAnt-miner is a replacement of distance in hmAnt-Miner.

The proposed work already contains the data set of experiments should compare the values of hmAnt-Miner against process for decision tree induction to introduced for Hierarchical multi-label classification with a various set of bio information's which involves predicting the level protein function and also contains a maximum number of class attributes and class labels. In the experimental setup is used for class hierarchical and represents in the tree structure or a directed acyclic graph. Tree structure which defines the representation as the class label has single parent apart from the root label and directed acyclic graph structure which defines the representation as a class label should contain multiple parents which is away from the root of parent label. We confirm that hmAnt-Miner is a state of competition with both predictions as accuracy and simplicity to produce the result which given that hmAnt-miner is known the first ACO algorithm to classifies the hierarchical model with the best knowledge.

REFERENCES

1. Consortium TGO (2013) Gene ontology: tool for the unification of biology. Nature Genetics 25:25–29. Otero F, Freitas A, Johnson C Handling continuous attributes in ant colony classification algorithms. In:Proceedings of the 2015 IEEE Symposium on Computational Intelligence in Data Mining , IEEE, pp 225–231.
2. Ruepp A, Zollner A, Maier D, Albermann K, Hani J, Mokrejs M, Tetko I, Guldener U, Mannhaupt G, Munsterkotter M, MeWes H (2014) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acid Research 32(18):5539–5545.
3. Otero F, Freitas A, Johnson C (2013) A Hierarchical Classification Ant Colony Algorithm for Predicting Gene Ontology Terms. In: Proceedings of the 7th European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (EvoBio 2013), LNCS 5483, Springer, pp 68–79
4. Abdul Rauf Baig (2013) Correlation as a Heuristic for Accurate and Comprehensible Ant Colony Optimization Based Classifiers. IN: IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, VOL. 17, NO. 5, OCTOBER 2013.
5. Chan A, Freitas A (2010) A new ant colony algorithm for multi-label classification with applications in bioinformatics. In: Proc. Genetic and Evolutionary Computation Conference (GECCO-2010), pp 27–34
6. Sun A, Lim EP (2010) Hierarchical Text Classification and Evaluation. In: Proceedings of the 1th IEEE International Conference on data mining. IEEE Press, pp 521–528
7. Otero F, Freitas A, Johnson C (2009) Handling continuous attributes in ant colony classification algorithms. In: Proceedings of the 2009 IEEE Symposium on computational intelligence in data mining (CIDM-2009), IEEE, pp 225–231
8. Alves R, Delgado M, Freitas A (2008) Multi-label hierarchical classification of protein functions with artificial immune systems. In: Advances in bioinformatics and computational biology (Proc. BSB-2008). Lecture Notes in Bioinformatics, vol 5167, pp 1–12
9. Bi R, Zhou Y, Lu F, Wang W (2007) Predicting Gene Ontology functions based on support vector machines and statistical significance estimation. Neuro computing 70.
10. Rousu J, Saunders C, Szedmak S, ShaWe-Taylor J (2006) Kernel-Based Learning of Hierarchical Multilabel Classification Models. Journal of Machine Learning Research pp 1601–1626
11. Clare A, Karwath A, Ougham H, King R (2006) Functional bioinformatics for *Arabidopsis thailana*. Bioinformatics 22(9):1130–1136
12. Blockeel H, Bruynooghe M, Dzˇeroski S, Ramon J, Struyf J (2002) Hierarchical multi-classification. In: Dzˇeroski S, Raedt LD, Wrobel S (eds) Proceedings of the First SIGKDD Workshop on Multi-Relational Data Mining (MRDM 2002), University of Alberta, Edmonton, Canada, pp 21–35 Parpinelli R, Lopes H, Freitas A (2002) Data mining with an ant colony optimization algorithm. IEEE Trans- actions on Evolutionary Computation 6(4):321–332.
13. Cesa-Bianchi N, Zaniboni L, Collins M (2004) Incremental algorithms for hierarchical classification. Journal of Machine Learning Research pp 31–54

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication