IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A Survey: Web Log Mining using Genetic Algorithm

Ranu Singhal[*1], NirupamaTiwari[2]

[*1,2] Shri Ram College of Engineering & Management, India

Abstract

Web mining has become a vast area of Research in last few years. Web Mining Which deals with the extraction of interesting knowledge from logging information produced by web server. In this paper we present a survey generate clusters using a multi objective genetic algorithm. Genetic algorithm is also a very hot area of research. In this paper we will compare the error value between FCM(Fuzzy c-means) and FCM-MOGA(Fuzzy C-Means multi objective Genetic algorithm). Genetic algorithm follows some steps and produce optimize solution. In using GA standard deviation and iteration value also affected. In this paper we have survey various paper based on Web log mining and FCM and GA.

Keywords: Web log mining,Fuzzy C-means, Genetic algorithm, clustering.

Introduction

Web Usage Mining is that part of Web Mining which deals with the extraction of knowledge from server log files; source data mainly consist of the (textual) logs that are collected when users access Web servers and might be represented in standard formats i.e. common log format. Extended log format. Log ML.

Clustering is a technique of dividing data into several clusters (groups or segments) where each cluster can be assigned several members together. One of the partitioning clustering techniques is K-means, which partitions the data in the form of two or more clusters or groups [2].

One of the methods to improve the performance of the KMeans clustering is the genetic algorithm. An application of genetic algorithm in optimization of K-Means clustering, among others, is in the search for images based on color feature with a GA-K-Means Clustering [3].

Fuzzy C-Means

The World Wide Web has huge amount information [4,5]and large datasets are available in databases. Information retrieving on websites is one of possible ways how to extract information from these datasets is to find different clusters of similar units for the description of the data vector descriptions are usuallyused each its component corresponds to a variable which can be measured in different scales (nominal, ordinal, or numeric) most of the well known clustering methods are implantedonly for numeric data(k-means method) or are too complex for clustering large datasets(such as hierarchical methods based on dissimilarity matrices). Fuzzy clustering[6] relevant for information retrieval as a document might be relevant to multiple queries ,Fuzzy clustering seems a natural technique for document categorization there are two basic methods of fuzzy clustering[4] ,one which is based on fuzzy c-partitions is called a fuzzy c-means clustering method and the other ,based on the fuzzy equivalence relations is called a fuzzy equivalence clustering method.

Clustering

Clustering is the process of grouping objects together in such a way that the objects belonging to the same group are similar and those belonging to different groups are dissimilar. Clustering [7] technique can be used in many applications for example biological, financial applications and many more. One of these application types is Web clustering where different types of objects can be clustered into different groups for various purpose.Clustering is a unsupervised method. This is follow a iteration process. Clustering have a different types. Like as partition clustering, density clustering, hierarchical clustering, Grid clustering. But in our survey partition clustering is used.

Partition clustering :

Partition Clustering technique create a one level partitioning of the data point. If K is the desired number of clusters, the partitioned approaches typically find all K cluster at once . This Technique based on the idea that a center point can represent a cluster

Genetic Algorithm

The concept of GA was developed by Holland and hiscolleagues in the 1960s and 1970s**.**Genetic Algorithm generate a population of point of each iteration. The best point in the population approaches an optimal solution. After that select the next population by computation which uses random number generators.genetic algorithm has work only single function.Genetic Algorithm use AI Search technique and produce a heuristic solution.

GA operates with a collection of chromosomes, called a population. The population is normally randomly initialized.Holland also presented a proof of convergence (the schema theorem) to the global optimum where chromosomes are binary. GA use two operators to generate new solutions from existing ones: crossover and mutation. The crossover operator is the most important operator of GA. In crossover, generally two chromosomes, called parents, are combined together to form new chromosomes, called offspring. The parents are selected among existing chromosomes in the population with preference towards fitness so that offspring is expected to inherit good genes which make the parents fitter. By iteratively applying the crossover operator, genes of good chromosomes are expected to appear more frequently in the population, eventually leading to convergence to an overall good solution. The mutation operator introduces random changes into characteristics of chromosomes. Mutation is generally applied at the gene level. In typical GA implementations, the mutation rate (probability of changing the properties of a gene) is very small and depends on the length of the chromosome. Therefore, the new chromosome produced by mutation will not be very different from the original one. Mutation plays a critical role in GA.Traditional GA are customized to accommodate multi-objective
problems by using specialized fitness functions and introducing methods to promote solution diversity.

Multi Object Genetic Algorithm

GA is based on population based approach. And it is easily solving multi objective optimization problem.GA can be modified to find a set of multiple non-dominated solutions in a single run. The ability of GA to simultaneously search different regions of a solution space makes it possible to find a diverse set of solutions for difficult problems with non-convex, discontinuous, and multi-modal solutions spaces. The crossover operator of GA may exploit structures of good solutions with respect to different objectives to create new non dominated solutions in unexplored parts of the Pareto front. In addition, most multi-objective GA do not require the user to prioritize, scale, or weigh objectives. Therefore GA have been the most popular heuristic approach to multi-objective design and optimization problems.

1.4.1 General approaches to multi objective optimization[8]:

1) One is to combine the individual objective functions into a single composite function or move all but one objective to the constraint set. In the former case, determination of a single objective is possible with methods such as utility theory, weighted sum method, etc.

2) The second general approach is to determine an entire Pareto optimal solution set or a representative subset. A Pareto optimal set is a set of solutions that are Non dominated with respect to each other. While moving from one Pareto solution to another, there is always a
certain amount of sacrifice in one objective(s) to achieve a certain amount of gain in the other(s). Pareto optimal solution sets are often preferred to single solutions because they can be practical when considering real-life problems.

In both cases, an optimization method would return a single solution rather than a set of solutions that can be examined for trade-offs. For this reason, decision-makers often prefer a set of good solutions considering the multiple objectives

Various Steps of Genetic Algorithm

1. Population
2 Selection
3 crossover
4 Mutation
5 Result

Population : This is a initial step of genetic Algorithm. In this step we have to collect data from data source .and after that design a set of values.And define the size of population. Like as Set S ={1,2,3----N} in this set size of population is N.

Selection :. In Selection procedure we use a pareto Rank Based Method. This method we can not use external population. Pareto rank method parallel handle error rate and iteration rate.

Pareto Ranking Approach

Pareto-ranking approaches[8] explicitly utilize the concept of Pareto dominance in evaluating fitness or assigning selection probability to solutions. The population is ranked according to a dominance rule, and then each solution is assigned a fitness value based on its rank in the population, not its actual objective function value. Note that here in all objectives are assumed to be minimized. Therefore, a low rank corresponds to a better solution in the following discussion.

Step1: Start with a random initial populationP0. Set t = 0.

Step2: If the stopping criterion is satisfied, return P_t.
Step3: Evaluate fitness of the population as follows:
Step3.1: Assign a rank r(x,t) to each solution xϵP_t using the ranking scheme given in Eq. (2).
Step3.2: Assign a fitness values to each solution based on the solution's rank as follows[36]:

$f(x,t) = N - \sum_{k=1}^{r(x,t)-1} n_k - .5 * (n_{r(x,t)} - 1)$

where n_k is the number of the solutions with rank k.
Step3.3: Calculate the niche count nc(k,t) of each solution x ϵ P_t using Eq. (4).
Step3.4: Calculate the shared fitness value of each solution x ϵ P_t as follows:
F'(x,t) = f(x,t)/nc(x,t).
Step3.5: Normalize the fitness values by using the shared fitness values

$F''(x,t) = \frac{f'(x,t) \, n_{r(x,t)}}{\sum_{y \epsilon p_t} f'(x,t)}$ f(x,t)

Step4: Use a stochastic selection method based onf'' to select parents for the mating pool. Apply crossover andmutation on the mating pool until offspring populationQ_t of size N is filled. Set $P_{t+1} = Q_t$
Step5: Set t=t+1, go to Step 2

Crossover : After we have decided what encoding we will use, we can make a step to crossover. Crossover selects genes from parent chromosomes and creates a new offspring. The simplest way how to do this is to choose randomly some crossover point and everything before this point copy from a first parent and then everything after a crossover point copy from the second parent.

Mutation : After a crossover is performed, mutation take place. This is to prevent falling all solutions in population into a local optimum of solved problem. Mutation changes randomly the new offspring. For binary encoding we can switch a few randomly chosen bits from 1 to 0 or from 0 to 1. Mutation can then be following

Result : This is a final step of genetic algorithm. In this step we got a results in cluster form.

Application

1. It is access a university data quickly. And automatically create access model.
2. Data Optimization
3. Robotics

Advantage and Disadvantage

The main advantage of the weighted sum approach is a straightforward implementation. Since a single objective is used in fitness assignment, a single objective GA can be used with minimum modifications. In addition, this approach is computationally efficient. The main disadvantage of this approach is that not all Pareto-optimal solutions can be investigated when the true Pareto front is non-convex. Therefore, multi-objective GA based on the weighed sum approach have difficulty in finding solutions uniformly distributed over a non-convex tradeoff surface.

Conclusion

The objective of this paper is present an overview of multiple-objective optimization methodsusinggenetic algorithms (GA).In this paper survey on FCM and Genetic Algorithm. In this paper we also describe clustering an entropy terminology. We also focus Genetic Algorithm procedure and how to select data and after that how to proceed in this paper. Our future work is reduce the error and iteration value.

References

[1] YandraArkeman, NursintaA.Wahanani, Aziz KustiyoBogor Agricultural University, Indonesia International Journal of Electrical & Computer Sciences IJECS-IJENS Vol:12 No:05
[2] Berkhin, Pavel. 2002. Survey of Clustering Data Mining Technique." http://www.ce.ucr.edu/barth/EE242/clustering_survey.pdf.
[3] Widodo, Yanu. 2008. PencarianGambarBerdasarkanFiturWarnaDengan GA-KMeans Clustering." JurusanTeknologiInformasi, PoliteknikElektronikaNegeri Surabaya.
[4] J. Srivastava, R. Cooley, M. Deshpande, PN. Tan, Web usage mining: discovery and applications of usage patterns from web data. SIGKDD Explorations, Vol. 1, No. 2, 2000, pp.12–23.
[5] M. N. Garofalakis, R. Rastogi, S. Seshadri, K. Shim Data minino and the web: past, present and future, In Proc. of the second international workshop on webinformation and data management, ACM, 1999.
[6] IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 1, January 2011 ISSN (Online): 1694-0814 www.IJCSI.org
[7] Proceedings of the 5th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, Madrid, Spain, February 15-17, 2006 (pp237-242
[8] Abdullah Konaka,, David W. Coitb, Alice E. SmithcaInformation Sciences and Technology, Penn State Berks, USA