ISSN (Online) 2249 - 054 X

International Journal of Computing and Corporate Research

Specialized and Refereed Journal for Research Scholars, Academicians, Engineers and Scientists

http://www.ijccr.com

International Manuscript ID : ISSN2249054X-V3I3M6-052013

VOLUME 3 ISSUE 3 May 2013

AN EFFECTIVE ALGORITHMIC APPROACH FOR FITNESS FUNCTION IN CLUSTERING

[1]*Payal Joshi* [2] *Arvind Selwal* [3]*Anuradha Sharma*

Department Of Computer Science Engineering

Ambala College of Engineering & Applied Research, Devasthali, Ambala-133101

Ambala City, India

ABSTRACT

The calculation and approximation of the parameters has been a growing thrust area in multiple domains including data mining, machine learning, neural networks, web mining, cloud computing and many others. The fitness function is used to associate the metric with related content or parameters. fitness function is the specific class of objective function that is used to summarize, as a single figure of merit and how close a given design solution is to achieving the set goals. More specifically, in the assorted domains, each design solution is represented as a value or set or function.

ISSN (Online) 2249 - 054 X

International Journal of Computing and Corporate Research

Specialized and Refereed Journal for Research Scholars, Academicians, Engineers and Scientists

http://www.ijccr.com

International Manuscript ID : ISSN2249054X-V3I3M6-052013

VOLUME 3 ISSUE 3 May 2013

Each design solution, therefore, should to be awarded a value of merit, to indicate how close it came to meeting the overall requirements, and this is generated by applying the fitness function to the test, or simulation, results obtained from that solution. This paper proposes an empirical and novel methodology for calculation of fitness function. The fitness function shall make use of percentile based cumulative implementation of the occurrences of data items in the database relations. The existing techniques for cluster formation and fitness function lack percentile cumulative results. This manuscript attempts to implement the proposed methodology in terms of efficiency, accuracy and performance.

KEYWORDS: Clustering, Fitness Value, Outlier Detection

INTRODUCTION

Clustering is an important KDD technique with numerous applications, such as marketing and customer segmentation. Clustering typically groups data into sets in such a way that the intra-cluster similarity is maximized and while inter-cluster similarity is minimized. Clustering is an unsupervised learning. Clustering algorithms examines data to find groups of items that are similar[5]. Research community have proposed different clustering algorithm and many are suitable for clustering numerical data. In real world scenario, data in database are categorical in nature, which are raw or unsummarized data, where the attributes cannot be pre-arranged as numerical values. Clustering categorical data is a major challenge in data mining [6].

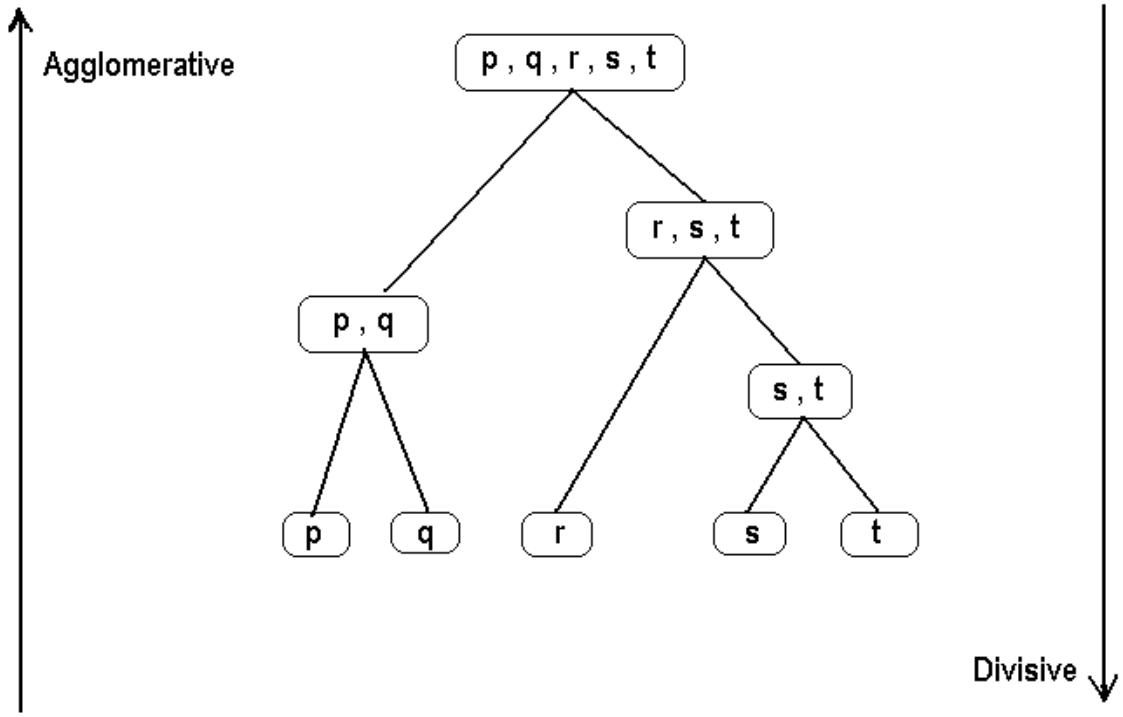
ISSN (Online) 2249 - 054 X
International Journal of Computing and Corporate Research
Specialized and Refereed Journal for Research Scholars, Academicians, Engineers and Scientists

http://www.ijccr.com

International Manuscript ID : ISSN2249054X-V3I3M6-052013

VOLUME 3 ISSUE 3 May 2013

An outlier in a dataset is defined informally as an observation that is considerably different from the remainders as if it is generated by a different mechanism. Mining for outliers is an important data mining research with numerous applications. Including credit card fraud detection, discovery of criminal activities in electronic commerce, weather prediction, marketing and customer segmentation [6].



LITERATURE SURVEY

ISSN (Online) 2249 - 054 X

International Journal of Computing and Corporate Research

Specialized and Refereed Journal for Research Scholars, Academicians, Engineers and Scientists

http://www.ijccr.com

International Manuscript ID : ISSN2249054X-V3I3M6-052013

VOLUME 3 ISSUE 3 May 2013

André Baresel et al [1] : Evolutionary Structural Testing uses Evolutionary Algorithms (EA) to search for specific test data that provide high structural coverage of the software under test. A necessary characteristic of evolutionary structural testing is that the fitness function is constructed on the basis of the software under test. The fitness function itself is not of interest for the problem; however, a well-constructed fitness function may substantially increase the chance of finding a solution and reaching higher coverage. Better guidance of the search can result in optimizations with less iterations, therefore leading to savings in resource expenditure. This paper presents research results on suggested modifications to the fitness function leading to the improvement of evolutionary testability by achieving higher coverage with less resources. A set of problems and their respective solutions are discussed.

M. Davarynejad et al [2] : Computational complexity is a major challenge in evolutionary algorithms due to their need for repeated fitness function evaluations. Here, we aim to reduce number of fitness function evaluations by the use of fitness granulation via an adaptive fuzzy similarity analysis. In the proposed algorithm, an individual's fitness is only computed if it has insufficient similarity to a queue of fuzzy granules whose fitness has already been computed. If an individual is sufficiently similar to a known fuzzy granule, then that granule's fitness is used instead as a crude estimate. Otherwise, that individual is added to the queue as a new fuzzy granule. The queue size as well as each granule's radius of influence is adaptive and will grow/shrink depending on the population fitness and the number of dissimilar granules. The proposed technique is applied to a set of 6 traditional optimization

ISSN (Online) 2249 - 054 X
International Journal of Computing and Corporate Research
Specialized and Refereed Journal for Research Scholars, Academicians, Engineers and Scientists

http://www.ijccr.com

International Manuscript ID : ISSN2249054X-V3I3M6-052013

VOLUME 3 ISSUE 3 May 2013

benchmarks that are for their various characteristics. In comparison with standard application of evolutionary algorithms, statistical analysis reveals that the proposed method will significantly decrease the number of fitness function evaluations while finding equally good or better solutions.

Andrew L. Nelson et al [3] : This manuscript highlights the study of fitness functions used in the stream and domain of robotics. This domain is the stream of research that applies artificial evolution to extract the control systems for autonomous robots. In this manuscript and research work, robots attempt to implement the task in a given environment using fitness function. The controllers in the enhanced performing robots are elected, tainted and propagated to execute the task again in an iterative progression that mimics some aspects of natural evolution. A key component of this process _ one might argue, the key component _ is the measurement of fitness in the evolving controllers. ER is one of a host of machine learning methods that rely on interaction with, and feedback from, a complex dynamic environment to drive synthesis of controllers for autonomous agents. These methods have the potential to lead to the development of robots that can adapt to uncharacterized environments and which may be able to perform tasks that human designers do not completely understand. In order to achieve this, issues regarding fitness evaluation must be addressed. In this paper we survey current ER research and focus on work that involved real robots. The surveyed research is organized according to the degree of a priori knowledge used to formulate the various fitness functions employed during evolution. The underlying motivation for this is to identify methods that allow the

ISSN (Online) 2249 - 054 X
International Journal of Computing and Corporate Research
Specialized and Refereed Journal for Research Scholars, Academicians, Engineers and Scientists

http://www.ijccr.com

International Manuscript ID : ISSN2249054X-V3I3M6-052013

VOLUME 3 ISSUE 3 May 2013

development of the greatest degree of novel control, while requiring the minimum amount of a priori task knowledge from the designer.

R. Ranjani et al [4] proposed Enhanced Squeezer algorithm, which incorporates Data-Intensive Similarity Measure for Categorical Data (DISC) in Squeezer Algorithm. DISC measure, cluster data by understanding domain of the dataset, thus clusters formed are not purely based on frequency distribution as many similarity measures do.

Payal Joshi et al [5] proposed a new approach for cluster formation and outlier detection. This approach makes use of fitness value. A fitness value is assignes to each tuple and based upon this fitness value the tuples are assigned to the clusters. If the difference between fitness value and threshold value is very large, the tuples are assigned to the outlier cluster.

Payal Joshi et al [6] proposed a new algorithm for clustring and outlier detection for categorical data. This algorithm has been developed with multiple parameters and assorted machine learning techniques. This algorithm makes use of multilayered approach for clustering and classification of the data sets.

PROPOSED APPROACH

In the algorithm given below, for calculation of fitness value our first step is to count the occurences of each product id. Then upper bound is calculated based on the

ISSN (Online) 2249 - 054 X
International Journal of Computing and Corporate Research
Specialized and Refereed Journal for Research Scholars, Academicians, Engineers and Scientists

http://www.ijccr.com

International Manuscript ID : ISSN2249054X-V3I3M6-052013

VOLUME 3 ISSUE 3 May 2013

percentage of the occurence of each product. Here, TopPercentage is set as upper bound. Further, percentile is calculated of each product based on the upper bound. In the next step of algorithm, clusters are formed with respect to percentile. Thereafter, if difference of threshold and percentile is same, select cluster arbitrarily otherwise tuple goes to the cluster for which difference is minimum. Then, get statistics and performance report.

PROPOSED ALGORITHM/PSEUDOCODE

1. Count occurrences of each product id/parameters/interest factor

2. Calculation of Upper Bound based on the Percentage of the occurrence

TopPercentage=UB

3. Calculate Percentile of each product/parameter based on the UB

4. Cluster Formation w.r.t percentile

5. If difference is same, select cluster arbitrarily

6. Get statistics and performance report

ISSN (Online) 2249 - 054 X

International Journal of Computing and Corporate Research

Specialized and Refereed Journal for Research Scholars, Academicians, Engineers and Scientists

http://www.ijccr.com

International Manuscript ID : ISSN2249054X-V3I3M6-052013

VOLUME 3 ISSUE 3 May 2013

Implementation of Dynamic Cluster Formation using Fitness Function and Proportional Analysis with the Threshold

Product ID	Price	WeekDay	Date	Month	Year	Items
001	950	wednesday	12	january	2013	89
001	950	sunday	23	march	2013	9
002	1020	saturday	12	march	2013	5
001	950	sunday	30	march	2013	8
002	1020	saturday	19	march	2013	33
003	2400	monday	3	march	2013	21
009	8000	saturday	19	february	2013	12
001	950	monday	11	march	2013	400
001	950	monday	11	march	2013	1200

Cumulative Units / Items Sold : 1777

Aggregation and Analysis of the Sold Items based on the ProductType

Occurence of : Product ID (001) | Count 1706 | Percentage is 96.
Occurence of : Product ID (002) | Count 38 | Percentage is 2.1

ISSN (Online) 2249 - 054 X

International Journal of Computing and Corporate Research

Specialized and Refereed Journal for Research Scholars, Academicians, Engineers and Scientists

http://www.ijccr.com

International Manuscript ID : ISSN2249054X-V3I3M6-052013

VOLUME 3 ISSUE 3 May 2013

Occurence of : Product ID (003) | Count 21 | Percentage is 1.1
Occurence of : Product ID (009) | Count 12 | Percentage is 0.6
Upper Bound : 96
Lower Bound : 1

Product ID (001) | | Percentile : 100
Product ID (002) | | Percentile : 2.2
Product ID (003) | | Percentile : 1.2
Product ID (009) | | Percentile : 0.7

Aggregation and Analysis of the Sold Items based on the Week Days

On : (monday) | Count 1621 | Percentage 91.22
On : (wednesday) | Count 89 | Percentage 5.008
On : (saturday) | Count 50 | Percentage 2.813
On : (sunday) | Count 17 | Percentage 0.956

CONCLUSION

In this paper an algorithm for the calculation of fitness function is proposed. This algorithm uses percentile based method for calculating fitness value which existing

ISSN (Online) 2249 - 054 X

International Journal of Computing and Corporate Research

Specialized and Refereed Journal for Research Scholars, Academicians, Engineers and Scientists

http://www.ijccr.com

International Manuscript ID : ISSN2249054X-V3I3M6-052013

VOLUME 3 ISSUE 3 May 2013

techniques lacks. The proposed and proved algorithmic approach using a web based simulator analyze various factors and parameters depending upon the run time fitness function value and assigns the weight or relevance. The proposed algorithm gives optimized results in terms of complexity as well as execution time on multiple running scenarios.

REFERENCES

[1] Andre baresel, harmen Sthamer, Michael Schmidt,2002. Fitness Function Design to improve Evolutionary Structural Testing

[2] M.Davarynejad, M.-R.Akbarzadeh-T, N.Pariz,2007. A Novel Framework for Evolutionary Optimization: Adaptive Fuzzy Fitness Granulation

[3] Andrew L.Nelson, Gregory J.Barlow, Lefteris Doitsidis,2008 .Fitness Functions in Evolutionary Robotics: A Survey and Analysis

[4] R.Ranjini, S.Anitha Elavarasi, J.Akilandeswari,2012. Categorical Data Clustering Using Cosine Based Similarity for Enhancing the Accuracy of Squeezer Algorithm

[5] Payal Joshi, Arvind Selwal, Anuradha Sharma,2013. An Effective Algorithmic Approach for Clustering and Boundary Analysis in Heterogenous Database Applications

ISSN (Online) 2249 - 054 X

International Journal of Computing and Corporate Research

Specialized and Refereed Journal for
Research Scholars, Academicians, Engineers and Scientists

http://www.ijccr.com

International Manuscript ID : ISSN2249054X-V3I3M6-052013

VOLUME 3 ISSUE 3 May 2013

[6] Payal Joshi, Arvind Selwal, 2013. The Scintillation of Algorithmic Approach in Fitness Based Clustering and Outlier Analysis