Turkish Journal of Computer and Mathematics Education *Vol.12 No.3(2021), 5051-5062*

Research Article

A Novel Granularity Optimal Feature Selection based on Multi-Variant Clustering for High Dimensional Data

Srinivas Kolli[a], M. Sreedevi[b]

[a] Research Scholar,[b] Professor,
[a, b]Department of Computer Science and Engineering, KoneruLakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India
[a]kollisreenivas@gmail.com, [b] msreedevi_27@kluniversity.in

Article History: Received: 10 November 2020; Revised 12 January 2021 Accepted: 27 January 2021; Published online: 5 April 2021

Abstract: Clustering is the most complex in multi/high dimensional data because of sub feature selection from overall features present in categorical data sources. Sub set feature be the aggressive approach to decrease feature dimensionality in mining of data, identification of patterns. Main aim behind selection of feature with respect to selection of optimal feature and decrease the redundancy. In-order to compute with redundant/irrelevant features in high dimensional sample data exploration based on feature selection calculation with data granular described in this document. Propose aNovel Granular Feature Multi-variant Clustering based Genetic Algorithm (NGFMCGA) model to evaluate the performance results in this implementation. This model main consists two phases, in first phase, based on theoretic graph grouping procedure divide features into different clusters, in second phase, select strongly representative related feature from each cluster with respect to matching of subset of features. Features present in this concept are independent because of features select from different clusters, proposed approach clustering have high probability in processing and increasing the quality of independent and useful features.Optimal subset feature selection improves accuracy of clustering and feature classification, performance of proposed approach describes better accuracy with respect to optimal subset selection is applied on publicly related data sets and it is compared with traditional supervised evolutionary approaches

Keywords: Clustering, Feature Selection, Genetic Algorithm, Granular Information, Multi-variant Calculation

1. Introduction

In recent years, most of the different data retrieval related applications such as gene related text data, data/text categorization, and retrieval of images with respect to different attributes/instance of features. Because of novel development of data technology high amount of data being collected, normally data itself it doesn't have reliable knowledge. Mining of data plays main role to mine similar patterns from large data sources, it is the main challenge for different mining related approaches to identify both relevant and irrelevant data from data source applications (which have different types of attributes/instances). Because of different/thousands of attributes present in large data sources, Feature selection (FS) is the basic data processing procedure to handle data mining efficiently, main aim behind features selection is to identify optimal feature subset which contain strongly relative and mostly matched data for making efficient decisions, feature selection may works with better comprehensive of a specific domain if the features are with good ability references. Because of inherent similarities in features, feature selection helps to reduce curse of dimensionality, optimal feature selection maximize or minimize the major importance of feature present in data sources. For recent data mining applications, different feature related selection approaches have been introduced for the studied of machine learning methods, these methods are classified into four different methods: Hybrid and Filter, Wrapper, Embedded methodologies.

Mostly embedded methodologies incorporate the selection of feature is sub part of training process to explore the features relates to specific domain related machine learning algorithms, and also these methods are most efficient than remaining approaches. Artificial neural networks, decision tree related approaches are most relevant to embedded categories, Wrapper related methods are used to predict accuracy of pre-described training calculation to determine effectiveness of selected subset features, accuracy usually high in categorical of different features. Filtered methods are used to learn independent of different learning calculations with good generality in features extraction, in these methods complexity is low with respect to low accuracy. Hybrid related methods are combination of filtered and wrapper methods to reduce search space in identifying sub-sequent wrapper conditions.

Feature selection approaches and cluster analysis has been implemented traditionally, in cluster related analysis, theoretic approaches are studied with different real time applications, also studied the feature selection

problem from perspective of feature granulation or sample granulation. For granulation of feature selection with respect to optimization genetic algorithm is proposed. Because of global search of features, genetic algorithm is used, in this representation each feature evaluated chromosome in GA represented as perspective feature space and stored in feature set then size of feature set described as size of granularity. Granularity based feature selection procedure described in figure 1.

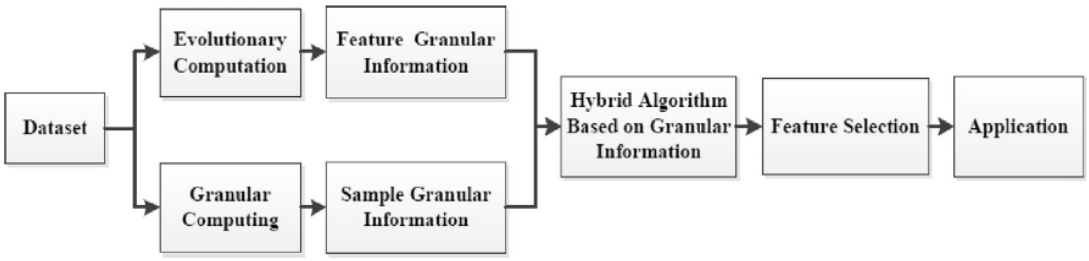

Figure 1. Granularity based feature selection procedure.

For sample data granulation Granularity Neighborhood relates rough-setis used. Based on all these procedures, propose and introduce a Novel Granular Feature Multi-variant Clustering based Genetic Algorithm (NGFMCGA) model. This model worked based on two steps. In first phase, based on theoretic graph grouping procedure divide features into different clusters, in second phase, select strongly representative related feature from each cluster with respect to matching of subset of features. Attributes present in different clusters have different relations i.e. independent, clustering strategy in NGFMCGA have highest probability in processing of relative and useful independent features. The proposed NGFMCGA approach was tested on publicly available text related datasets, experimental results of NGFMCGA show that, compare with other feature selection related algorithms/approaches, proposed approach only performs optimal features with improves the performance of well know attributes.

2. Review of Related Work

In this section, we present and discuss about different authors opinion regarding feature selection from different data sources. Basic traditional procedure relates to traditional approaches described as follows:

Feature subset determination can be seen as the procedure of distinguishing and evacuating the same number of immaterial and excess includes as could be expected under the circumstances. This is on the grounds that 1) insignificant highlights don't add to the prescient precision, also, 2) excess features don't redound to getting a better indicator for that they give for the most part data which is as of now present in different feature(s). Of the many component subset determination calculations, a few can adequately dispense with unessential features however neglect to handle excess features however some of others can dispose of the immaterial while taking care of the excess features. Traditionally proposed **F**ast clustering b**a**sed feature **S**election algori**t**hm (FAST) calculation falls into the subsequent gathering.

Generally, feature subset choice research has concentrated on looking for important feature. A notable model is Relief et.al, which gauges each component agreeing to its capacity to separate examples under various targets dependent on separation based measures work. Nonetheless, Help is insufficient at expelling excess highlights as two prescient however exceptionally connected features are likely both to be profoundly weighted. Help F broadens Relief, empowering this strategy to work with uproarious and inadequate informational indexes and to manage multi-class issues, yet at the same time can't recognize repetitive features.

3. Basic Preliminaries

Basic preliminaries used in this research described in this section

Genetic Algorithm (GA)

It is a well-known approach for potential global search with different features selection to different researchers, for efficient feature optimization basic procedure relates to genetic algorithm described as follows:

Let us consider the finite set of string i.e. $X=\left(x_1,x_2,.....,x_n\right)$ and $X\Rightarrow A=\left(a_1,a_2,.....,a_L\right)^T$ X be the encoding function of A represented in $e\left(A\right)$ whenever A be the decoding function of X represented as $A=e^{-1}(X)$. x_i be the gene expression and floating point binary representation of x_i , here length of encoding L, then $H_L=\left\{X=x_1,x_2,.....,x_L \mid x_i\in\Gamma,i=1,2,.....,L\right\}$ and it is demoted, for the individual feature selection

in population represented as H_p . For efficient feature optimization use search space procedure H_p and also calculates the fitness value of each and individual attribute, X satisfy $\max_{X \in H_p}(X)$.

Based on above procedure each attribute in record $GAF(e, J, S., E, \psi)$, where $e, J, S., E \& \psi$ represents as format of coding, metrics relates to fitness of each attribute, operator selection, parameter set and operator relates to genetic. Finally general steps of GA is

a) Generate N random individual features at initial data, evaluate all the features
b) Calculate the fitness function for each feature based on elected feature
c) Employ the selection of features with respect to cross-over and mutation to process next
d) Check fitness function for each feature and the evaluate the optimal feature set relates to each individual selected attribute
e) Terminate from process with different scenarios.

Granularity Neighborhood relates Rough-Set

In information retrieval system, granularity information is the measurement of data evaluationbased on knowledge and data. Data granularity is smaller than ability is stronger related to knowledge and vice-versa, Let us consider data relates to classification and then formalize the decision system describes as $DS =< U, X, D >$. The sample data set $U = \{a_1, a_2, ..., a_n\}$ & $X = (x_1, x_2, ..., x_N)$ is the feature set which is obtained from sample data set then decision of feature selection

In information retrieval system, $S = (U, X)$, $R \subseteq X, \& U / R = \{R_1, R_2, ..., R_m\}$ then data knowledge granularity of R described as $GK(R) = \frac{1}{|U|^2} \sum_{i=1}^{m} |R_i|^2$ where $\sum_{i=1}^{m} |R_i|^2$ describes the equivalent relation of determined with $\bigcup_{i=1}^{m} R_i \times R_i$.

All the scenarios used in proposed implementation with efficient selection of features.

4. Proposed NGFMCGA Procedure Model for Optimization

Basic Definitions

Feature subset selection to be selected for the identification and redundant/ irrelevant features removal. Different features consists different co-related attribute relevancy with respect to relevancy between each attribute. All these are keeping in mind develop a Novel Granular Feature Multi-variant Clustering based Genetic Algorithm (NGFMCGA) effectively remove redundant with respect to irrelevant features and obtain good subset feature.

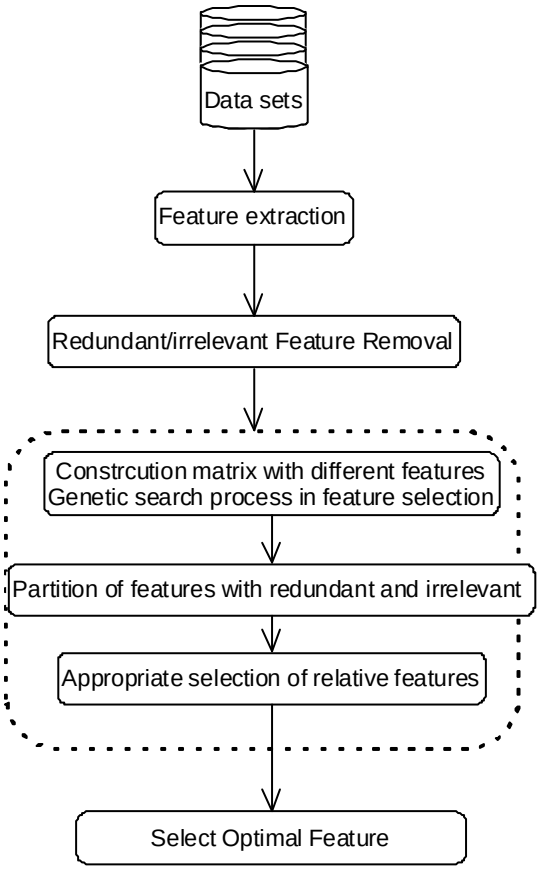

Figure 2. Feature selection procedure of the proposed approach

Feature selection procedure described in figure 2. It describes the removal of features relates to irrelevant and redundant, select appropriate relevant features to the selected feature by dismissing irrelevant features and also select different relevant features from representative clusters and then compute the final optimal feature subset.

Genetic with Granularity related Feature Selection Procedure

Feature selection with search space of different attributes evaluated with granularity based approach designed with measure of different features described as follows:

Let us consider the P be the population of all data sets, U be the different genes $U / P = \{S_1, S_2, ..., S_k\}$, where S_k be the selection of different features in last chromosome (attribute). Granularity data information $G(C_i)$ with combination of knowledge granularity $G_p(C)$ as

$$G(C_i) = -\frac{|S_i|}{|U|}\log_2\frac{|S_i|}{|U|}$$

$$G_p(C) = \frac{1}{K}\sum_{i=1}^{K} G(C_i)$$

Feature ID:	1	2	3	4	5	$G(C_i)$
Key value:	0	1	0	1	1	
			P_1			P_2

Figure 3. Representation of features in proposed approach**.**

It is to find the value of $G_p(C)$ is the size of overall feature set based on above conditions; let us consider the P be the different feature related characteristics where each feature selected from subset feature set shown in figure 3.

Threshold based Fitness Function

It is the process of selection of feature with respect to subset feature selection; fitness value is used to describe the accuracy of classification. Based on the sample distribution characteristics with respect to differences in class related features to measure adapted subset features. With in class representation, selected feature sub set ensures as small as possible, and between class labels are represented as long as possible

Let us consider $a_k^i \,\&\, a_k^j$ are the n-dimensional vector representation with respect to different feature classes i and j respectably, main representation of sample vector classes like i and j represented as $m_i = \frac{1}{n_i}\sum_{k=1}^{n_i} x_k^i$ and also all the sample vector classes represented as $m_i = \sum_{c=1}^{c} p_i m_i$, between class label differences $D_j = \sum_{i=1} p_i (m_i - m)(m_i - m)^T$ represented with-in class differences represented as $D_n = \sum_{i=1}^{c} p_i \frac{1}{n_i} \sum_{k=1}^{n_i} \left((x_k^i - m_i)(x_k^i - m_i)\right)^T$ where p_i be the corresponding class prior probability, to increase the performance of classification relates to selected features with respect to granularity information $G_p(C)$.

$$R(T \mid P_0) = 1 - \left(\left| \frac{F_M - F_1}{F_M - F_0} \right| \times \left| \frac{F_M - F_2}{F_M - F_1} \right| \times \ldots\ldots\ldots \times \left| \frac{F_M - F_T}{F_M - F_{T-1}} \right| \right)^{\frac{1}{T}}$$

$$1 - \left(\left| \frac{F_M - F_T}{F_M - F_0} \right| \right)^{\frac{1}{T}}$$

Let us consider P_0 be the population of initial, then convergence rate of genetic algorithm with different feature representations.

Algorithm 1 Optimal feature selection (OFS) procedure with respect to genetic algorithm features.

Input data: Let us consider i/p data is $D = (d_1, d_2, \ldots\ldots, d_n, y)$
$P_{size}, H_{crom\_len}, P_c, P_m, T$

Output data: Best Optimal feature sub set.

1. Initialise all the input parameters
2. Loading input data set
3. Normalize data // $v_n = \frac{(v - v_{\min})}{(v_{\max} - v_{\min})}, S_n = S(v_n)$
4. $for(i = 1 to P_{size}) do$
5. $for(i = 1 to H_{size}) do$
6. Randomly initialise the parameters $Pop(i, j) = round(rand)$
7. End for, end for
8. $for(i = 1 to P_{size}) do$
9. Then $Subset = S_n(:, find(Pop(i,:) == 1))$
10. End for and initialise iterative parameter i.e. t=0,

11. Execute While up to not meet T
12. $for(i = 1 to P_{size})$
13. Calculate individual fitness $Fit_{value}(i) = fit1(pop(i))$
14. $Weight\_Rank\left(Fit_{value}(i)\right)$
15. $Pop = Elit(pop)$
16. $Crome(t, Pop) = Chrom(t, Pop_c)$
17. $Crome(t, Pop_c) = Chrom(t, Pop_m)$
18. $New_{Pop} = Pop$
19. $t = t + 1$, While_ends, for t=1 to T
20. $Sl(t) = R(T \mid P_0)$, end for
21. Return $OFS = S(:, Find(Best_{indi}(1,:) == 1))$
22. Select best optimal feature from overall data set i.e. OFS

For each attribute fitness difference is represented with normalized geometric mean with respect to rate of convergence, so it is clear faster convergence defines larger convergence rate to FT and FM and then convergence rate is reach to maximum 1.

Based on above discussion, feature selection could be done with granularity based feature related genetic algorithm procedure discussed in algorithm 1, algorithm 1 describes initialization process with different input parameters listed from 4-9, at the same time fit () is used to evaluate the fitness function of individual parameters listed in 9-13, after feature extraction evaluate the optimal feature selection listed from 13-22 shown in algorithm 1. Generally because of uncertain property of genetic in our proposed algorithm improved to evaluate the time complexity issue based on decreasing the iterations using different parameters like no. of individual P_{size} , rate of mutation P_m, rate of crossover P_c, length of the individuals L. Optimized time complexity is described as $O\left(T \times L \times P_{size}^2\right)$.

Optimization of Granularity Feature Selection based on Neighborhood Genetic Calculation

In this section, for granularity feature selection, granularity radius is usedto explore different attributes, optimal feature set selection related data sets to be explored as input data from fitness function. Classification related pattern feature selection with respect to interpreted feature subspace, feature granularity can improve the reduction of dimensionality with extent features without selection of sample granularity. Granularity feature optimization procedure described in algorithm 2

Algorithm 2 Procedure to select granularity optimal features subset selection

I/P: Optimal set i.e. $OFS = \left(x_1, x_2, ..., x_n, y\right)$, feature sub set i.e. $S = \left(x_1, x_2, ..., x_n, y\right)$, no.of iterations (T), importance of Lower limit (χ),

O/P: Optimal granularity feature subset λ , feature subset S_F

1. Initialize all algorithm parameters and load input data S, OFS
2. For i=1 to n do For i=1 to P_r
3. For i=1 to m then For i=1 to H_r if (missing values exit) then
4. $x(i, j) = mean(OFS)\ \&\ Pop(i, j) = round$ //Random initialize of generate missing values
5. End if,for,for
6. If t==0 , while maximize the iterations for reduction set
7. For j=1 to m , For i=1 to P_r then $a_j = S_A - red\_set$
8. $\gamma_{red \cup a_j}\left(D^{\lambda}\right) = \dfrac{Card(N_{red \cup a_j}.D^{\lambda})}{Card(U)}$ //dependence calculation of each attribute a_i
9. $SIG\left(a_j, D^{\lambda}, red\right) = \gamma_{red \cup a_j}\left(D^{\lambda}\right) - \gamma_{red}\left(D^{\lambda}\right)$ //importance calculation for different parameters

10. $fitness\_value(i) = a + fitness\_value(i)^{*}(b-a)/(2^{Hr}-1)$ //
11. $\vec{\delta}_i = \frac{1}{\lambda_i}(\tau_1, \tau_1, ..., \tau_1)$, End For, For
12. $k-nnClassify(tr_{data}, tr_{attr\_label}, te_{data}, Eucliden, Random)$ //apply k-nn for attribute partition based on euclidean granularity radius for each attribute
13. For i=1 to n do $Rank(fit_{value}(i))$ $Fit_{value}(i) = fit2(pop(i))$
14. $t = t+1$, End While
15. Obtain best from individuals
16. Return optimal granularity based with feature sub set

According to the threshold fitness function, algorithm 2 describes the granularity optimal feature selection, from 1-4 line defines basic definations relates to ranularity feature selection, fit2() calculates the fitness for granularity optimal feature selection described from 6-13, from 14 calculates the time complexity of algorithm 2. Because of crossover, mutation fitness of individaul parameters and dimensionality reduction with different features, proposed approach performs efficient classification accuracy with optimal feature respctually. So that time complexity of optimized granularity feature algorithm describes as $O\left(T \times n \times m \times \left(n \log m + 1\right)\right)$ at different stages in processing of multiple attributes with optimal granularity features.

5. Experimental Evaluation

In this section, we describe the performance evalluation of proposed approach with comparison of different feature subset selection calculations/methods with different attributes in high dimensional data.

To analyze proposed approach performance from lowerative perspective selection of different sample related features selected randomly, data sets are downloaded from following UCI database repository link with different parameter sequences https://archive.ics.uci.edu/ml/datasets.php?format=&task=cla&att=&area=&numAtt=&numIns=&type=&sort=nameUp&view=tableAll these data describes sample attributes with high dimensions, low dimensions with sample attributes present in different real time statical analysis of company data sets described in table 1.

Table 1 Description of different data sets

S. No	Description of dataset	Different features	Different instances	Class references
1	WDBC	35	554	2
2	Sonar	66	235	2
3	Multi-Feature	650	2204	10
4	Stat_Synthetic	25-30	60-75	3-4
5	Leukemia	7318	76	3

Based on above data set description, present several types of experiments about proposed approach i.e. Novel Granular Feature Multi-variant Clustering based Genetic Algorithm (NGFMCGA) compared with FAST[1], FOCUS-SF [2], FCBF[3]. In order to increase the performance of proposed approachapplied on high dimensional data sets, classification accuracy behind selection of different features (|S|)and communication cost t(s), compare with traditional approaches and also improve precision, recall, memory utilization with respect to classifying features with time complexity in exploring correlation between different features for high dimensionality data sources.

Results

Classifying multi granularity features accuracy with different attribute processing values described in table 2.

Table 1 Accuracy of different data set values with different approaches.

Accuracy				
Dataset Description	Proposed Approach	FAST [1]	FOCUS-SF [2]	FCBF [3]
Data set 1	94.6	81.2	89.4	87.3
Data set 2	95.8	89.5	79.7	92.7
Data set 3	97.2	91.3	82.6	92.4
Data set 4	96.4	76.4	77.4	78.4
Data set 5	98.3	85.6	82.4	87.3

As shown in table 1, proposed approach describe the different accuracy values with exploring different features, when compared to traditional approaches, proposed approach gives better and efficient results with respect to processing multi-features with multi label data analysis which consists high dimensional attributes relations.

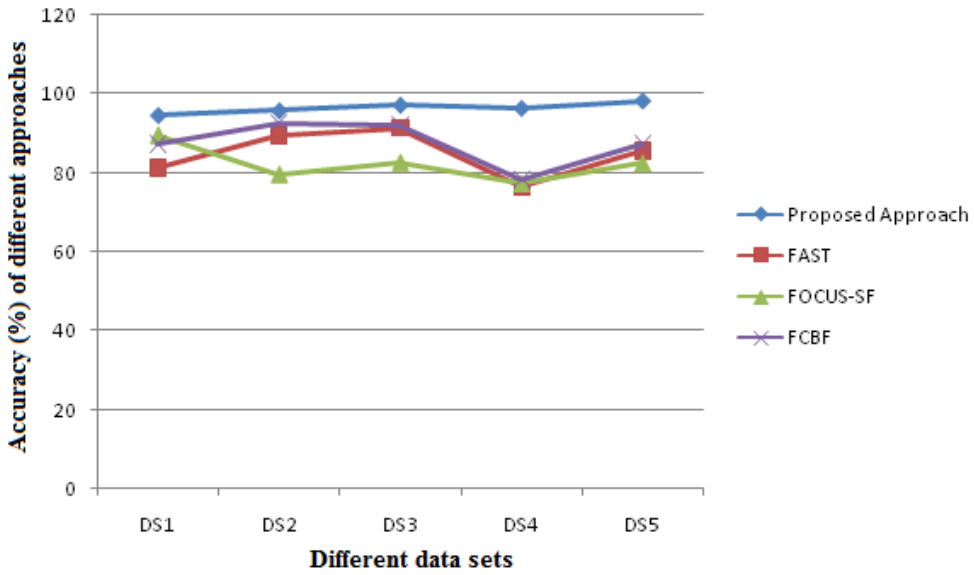

Figure 4 Performance evaluation of accuracy with comparison of different approaches

As shown in figure 4, it describes the accuracy of proposed approach with different labeled correlation between features for high dimensional data sources for optimal multi feature selection. Precision for granularity feature selection with different values processing described in table 3 with different attribute relations.

Table 3 Different precision values for different approaches

Precision				
Data sets Description	Proposed Approach	FAST [1]	FOCUS-SF [2]	FCBF [3]
Data set 1	0.56	0.352	0.45	0.37
Data set 2	0.52	0.321	0.43	0.42
Data set 3	0.572	0.40	0.31	0.38
Data set 4	0.545	0.43	0.392	0.37
Data set 5	0.64	0.353	0.34	0.35

As shown in table 3, precision values are described with different notations, performance of these notations described in figure 5. It describes the performance evaluation of precision with different attribute relations described as follows

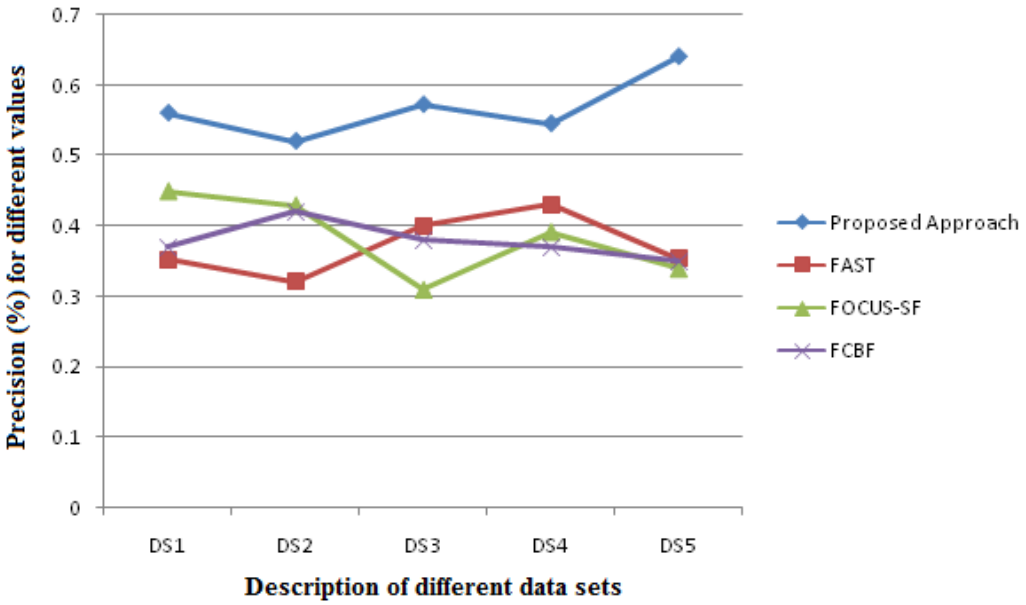

Figure 5. Performance evaluation of precision

As shown in figure 5, proposed approach gives better attribute partition i.e. above 0.5-0.7, remaining approaches were not reached the proposed approach precision and not increase more than 0.5, they are in only 0.3-0.5 only for different multi-dimensional data sets. Recall values for granularity feature exploration described in table 4.

Table 4. Granularity feature exploration recall values with different partitions.

Recall				
Databases	Proposed Approach	FAST [1]	FOCUS-SF [2]	FCBF [3]
Data set 1	0.71	0.541	0.61	0.51
Data set 2	0.65	0.48	0.50	0.54
Data set 3	0.69	0.51	0.46	0.62
Data set 4	0.71	0.57	0.42	0.41
Data set 5	0.72	0.42	0.41	0.39

Recall values of correctly identified optimal features relates to different data sets described table 4, all these values explore optimal feature from multi dimensional features with different relations. Performance evaluation described in figure 6.

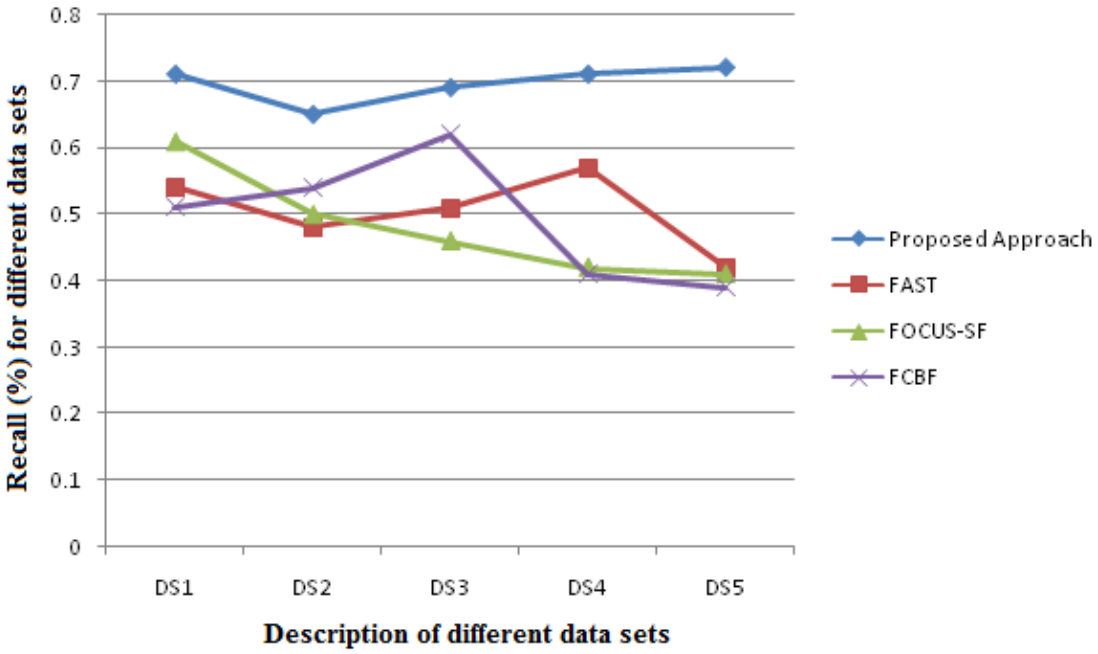

Figure 6. Performance of recall for different data sets with successive relations.

As shown in figure 6, it describes the recall values i.e. correctly identified values with different granularity features relations. Time complexity values for different algorithms described in figure 7.

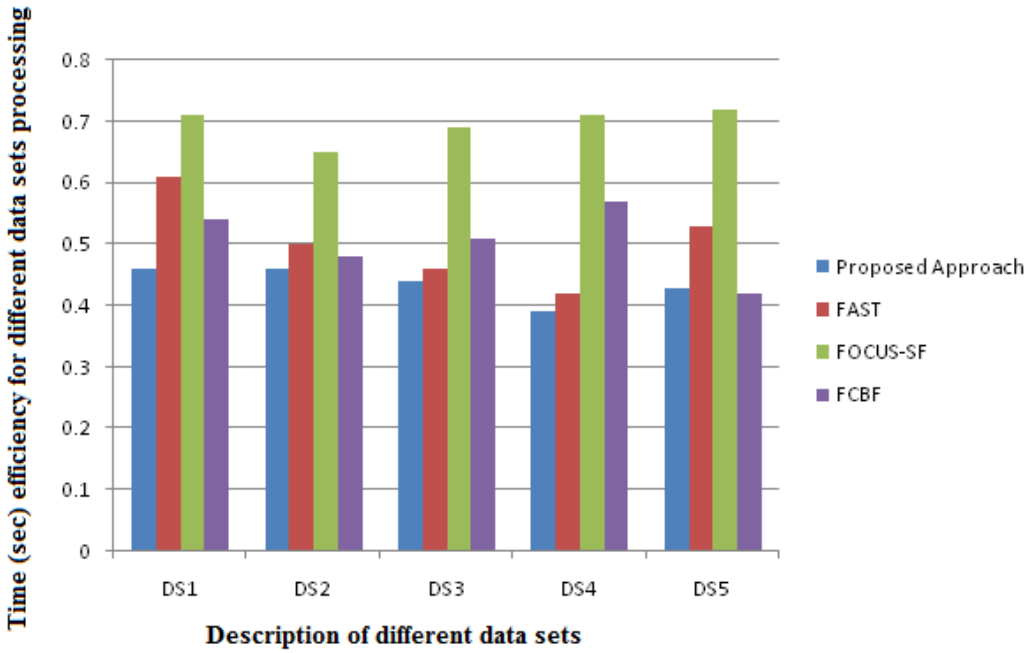

Figure 7 Performance of time efficiency with respect to different data sets

As shown in above figures proposed approach gives better optimal feature selection from overall multi dimensional data sets evaluation, In terms of time, accuracy, precision and recall for granularity multi-feature selection.

6. Conclusion

In this paper, granularity feature selection model i.e.Novel Granular Feature Multi-variant Clustering based Genetic Algorithm (NGFMCGA) is presented, this model mainly consists granulation of subset features based on genetic calculation and neighborhood sample granulation with multi features. For granularity feature optimization, granularity based genetic algorithm is presented with calculation of fitness and other sequence parameters and it also improve quality in selection of subset features. Proposed approach is also kept good offence on optimal subset feature selection and improves the classification accuracy in identification purely matched patterns for

synthetic related data sets. In the future work, we plan to explore new model to obtain multi-feature optimal selection with respect to multi-dimensions and multi-objective optimization in selection of features and improve the efficiency.

References

Qinbao Song, Jingjie Ni, and Guangtao Wang, "A Fast Clustering-Based FeatureSubset Selection Algorithm forHigh-Dimensional Data", proceedings in IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 1, JANUARY 2013.

H. Almuallim and T.G. Dietterich, "Learning Boolean Concepts in the Presence of Many Irrelevant Features," Artificial Intelligence,vol. 69, nos. 1/2, pp. 279-305, 1994.

L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," Proc. 20th Int'l Conf. Machine Leaning, vol. 20, no. 2, pp. 856-863, 2003.

Hongbin Dong, Tao Li∗, Rui Ding, Jing Sun, "A novel hybrid genetic algorithm with granular information forfeature selection and optimization", Applied Soft Computing 65 (2018) 33–46.

Alok Kumar Shukla*, Pradeep Singh† and Manu Vardhan, "A New Hybrid Feature Subset Selection Framework Based on Binary Genetic Algorithm and Information Theory" proceedings in International Journal of Computational Intelligence and Applications Vol. 18, No. 3 (2019) 1950020 (22 pages).

Shao-Lun Huang, Xiangxiang Xu," An Information-theoretic Approach to Unsupervised Feature Selection for High-Dimensional Data" proceedings in arXiv:1910.03196v1 [cs.IT] 8 Oct 2019.

SmitaChormungea,∗, Sudarson Jena, "Correlation based feature selection with clustering for highdimensional data", proceedings in Journal of Electrical Systems and Information Technology 5 (2018) 542–549.

G. J. Op't Veld and M. C. Gastpar, "Caching gaussians: Minimizing total correlation on the Gray-Wyner network," in 2016

Annual Conference on Information Science and Systems (CISS). IEEE, 2016, pp. 478–483.

S.-L. Huang, X. Xu, L. Zheng, and G. W. Wornell, "An information theoretic interpretation to deep neural networks," arXiv preprint arXiv:1905.06600, 2019.

E. Pashaei, N. Aydin, Binary black hole algorithm for feature selection andclassification on biological data, Appl. Soft Comput. (2017).

M. Seera, C.P. Lim, A hybrid intelligent system for medical data classification,Expert Syst. Appl. 41 (5) (2014) 2239–2249.

B. Xue, M. Zhang, W.N. Browne, X. Yao, A survey on evolutionary computationapproaches to feature selection, IEEE Trans. Evolut. Comput. 20 (4) (2016)606–626.

A. K. Shukla, P. Singh and M. Vardhan, A two-stage gene selection method for biomarkerdiscovery from microarray data for cancer classi¯cation, Chemometr. Intell. Lab. Syst.183 (2018) 47–58

J. R. Anaraki and H. Use¯, A feature selection based on perturbation theory, Expert Syst.

Appl. 127 (2019) 1–8.

A. K. Shukla, P. Singh and M. Vardhan, Dna gene expression analysis on di®use large bcelllymphoma (dlbcl) based on filter selection method with supervised classificationmethod, in Computational Intelligence in Data Mining (Springer, 2019), pp. 783–792.

R. Armañanzas, M. Iglesias, D. A. Morales and L. Alonso-Nanclares, Voxel-based diagnosisof alzheimer's disease using classi¯er ensembles, IEEE J. Biomed. Health Inf. 21(3)

(2017) 778–784.

J. Tang and S. Zhou, A new approach for feature selection from microarray data basedon mutual information, IEEE/ACM Trans. Comput. Biol. Bioinform. 13(6) (2016)1004–1015.

B. Liao, Y. Jiang, W. Liang, W. Zhu, L. Cai and Z. Cao, Gene selection usinglocality sensitive Laplacian score, IEEE/ACM Trans. Comput. Biol. Bioinf. 11(6) (2014)1146–1156.

N. Hoque, D. K. Bhattacharyya and J. K. Kalita, MIFS-ND: A mutual information-basedfeature selection method, Expert Syst. Appl. 41(14) (2014) 6371–6385.

M. Zalasiński, K. L/ apa and K. Cpałka, Prediction of values of the dynamic signaturefeatures, Expert Syst. Appl. 104 (2018) 86–96.

C. Yan, J. Ma, H. Luo and A. Patel, Hybrid binary coral reefs optimization algorithmwith simulated annealing for feature selection in high-dimensional biomedical datasets,Chemometr. Intell. Lab. Syst. 184 (2019) 102–111.

M. Ghosh, S. Begum, R. Sarkar, D. Chakraborty and U. Maulik, Recursivememetic algorithm for gene selection in microarray data, Expert Syst. Appl. 116 (2019)172–185.

R. Guha, M. Ghosh, S. Kapri, S. Shaw, S. Mutsuddi, V. Bhateja and R. Sarkar, Delugebased genetic algorithm for feature selection, Evol. Intell. (2019) 1–11.

W. Zhang, H. He and S. Zhang, A novel multi-stage hybrid model with enhanced multipopulationniche genetic algorithm: An application in credit scoring, Expert Syst. Appl.121 (2019) 221–232.

H. Li, X. Huang, P. Yang and H. Yang, A new pressure vessel design by analysis methodavoiding stress categorization, Int. J. Press. Vessels Pip. 152 (2017) 38–45

Srinivas Kolli, M. Sreedevi. 2018. Adaptive Clustering Approach to Handle Multi Similarity Indexfor Uncertain Categorical Data Streams. Jour of Adv Research in Dynamical & Control Systems, Vol. 10, 04- Special Issue, 2018.

Srinivas Kolli, M. Sreedevi A Novel Index based Procedure to Explore Similar Attribute Similarity in Uncertain Categorical Data, ARPN Journal of Engineering and Applied Sciences, Volume 14,Issue 12,2019

Srinivas Kolli, M. Sreedevi, PROTOTYPE ANALYSIS OF DIFFERENT DATA MINING CLASSIFICATION AND CLUSTERING APPROACHES, ARPN Journal of Engineering and Applied Sciences, Volume 13, Issue 09,2018

M.Sreedevi, Vijay Kumar, G. Valli Kumari,V,2014Parallel and distributed approach for incremental closed regular pattern mining, IEEE

Vijay Kumar, G. Valli Kumari, V. 2013. Incremental mining for regular frequent patterns in vertical format. International. 5(2): 1506-1511.

Vijay Kumar, G. Valli Kumari, V. 2012. Sliding window technique to mine regular frequent patterns in data streams using vertical format. IEEE International Conference on Computational Intelligence and Computing Research, ICCIC.

Vijay Kumar, G. Valli Kumari, V. 2012.Parallel and Distributed Frequent-Regular pattern mining using vertical format in large databases.IEEE