AN EFFECTIVE FEATURE SELECTION TECHNIQUE FOR MINING HIGH DIMENSIONAL DATA ON BIG DATA

BY

K. BHASKAR NAIK * S.P. SINDHUJA **

* Assistant Professor, Department of Computer Science and Engineering, Sree Vidyanikethan Engineering College, Tirupati, India.
** PG Scholar, Department of Computer Science and Engineering, Sree Vidyanikethan Engineering College, Tirupati, India.

ABSTRACT

In the recent years, many research innovations have come into foray in the area of big data analytics. Advanced analysis of the big data stream is bound to become a key area of data mining research as the number of applications requiring such processing increases. Big data sets are now collected in many fields eg., Finance, business, medical systems, internet and other scientific research. Data sets rapidly increase their size as they are often generated in the form of incoming stream. Feature selection has been used to lighten the processing load in inducing a data mining model, but mining a high dimensional data becomes a tough task due to its exponential growth of size. This paper aims to compare the two algorithms, namely Particle Swarm Optimization and FAST algorithm in the feature selection process. The proposed algorithm FAST is used in order to reduce the irrelevant and redundant data, while streaming high dimensional data which would further increase the analytical accuracy for a reasonable processing time.

Keywords: Feature Selection, Minimum Spanning Tree, Classification.

INTRODUCTION

Big Data corresponds to the huge volume of data. In the recent years, big data is becoming a crucial area of research. Big data are used in marketing sectors, banking sectors, medical systems, etc., for its immense use. In order to retrieve useful information from the huge volume of data, datamining methods are employed. The underlying concept in the datamining process is feature selection.

Feature selection refers to extraction of similar kinds of elements or particles or data. It is one of the critical step for various reasons [4]. Feature selection involves in the selection of predetermined number of features which leads to the performance enhancement of the total classifier. This technique is employed in various fields such as in text classification, speech recognition and medical diagnosis. But practically, there is a need to reduce the number of measurements without degrading the performance of the system [3].

The risk or complexity is high in feature selection because of complex interaction between the features. Another factor, i.e size of the search space is directly proportional to increase in complexity of feature selection process. The size increases in an exponential manner based on the number of features present in the dataset. There are different kinds of methods employed in the feature selection process with various dimension measures [2]. In the recent years, many strategies have come into foray to obtain optimal solutions for different problems, Evolutionary algorithms such as ant colony algorithm, genetic algorithm and Particle swarm optimization algorithm are used in optimization techniques.

Feature Selection algorithms have the capabilities to improve the performance in the classification process [10].

1. Background Analysis

1.1 Genetic Algorithms

In the Genetic Algorithms, a group of particles or a swarm of particle or candidate solutions is evolved in an optimization for obtaining better solution. Genetic algorithms are independent and can be applied to various problems irrespective of the domain area. Over

the years Genetic Algorithms have made a substantial improvement in the random as well as local search techniques, which also called as adaptive search techniques. The algorithm proceeds by collecting information about its neighborhood particles and forwards till it finds the best solution to the given search space [7]. One of the main disadvantages arousing in genetic algorithms for feature selection is the selection of evaluating function or fitness function. The efficiency and performance of the system depends upon the objective function which is chosen to evaluate the system. It forms the basis for the performance analysis.

1.2 Particle Swarm Optimization

Particle Swarm Optimization was first developed by Eberhart and Kennedy in 1995 [1] with a view of simulation of social system. This method has been tremendously utilized in various applications. Particle Swarm Optimization is an evolutionary computing technique. One of the major difference between evolutionary computing technique and Particle Swarm Optimization is, the former one uses genetic operators and the latter one uses physical movements of individuals in the swarm. One of the most beguiling features that provoke us to use PSO is its simple procedure and few parameters. PSO is similar to Genetic algorithm, for example it starts with neighborhood with randomly generated swarm, calculation of fitness value, evaluating, later updating of the position and search for the optimal solution, but it could not ensure the optimal solution. It consist of two best fit or best value. They are called Pbest and Gbest. Pbest corresponds to the best value that has been obtained so far and the Gbest corresponds to the best value it has achieved in the global space or the population. Pbest is also called local optimal solution and Gbest is called global optimal solution. The particle in the swarm are represented in N-Dimensional space and are characterized by their position and velocity [8]. The original or standard PSO is given as,

$$v_{id} = v_{id} + c_1 r_1 (p_{id} - x_{id}) + c_2 r_2 (p_{gd} - x_{id}) \quad (1)$$

$$x_{id} = x_{id} + v_{id} \quad (2)$$

The modified PSO [6] equation is given below.

$$v_{id} = w * v_{id} + c_1 r_1 (p_{id} - x_{id}) + c_2 r_2 (p_{gd} - x_{id}) \quad (3)$$

In the above equation, d represents the set of natural numbers ranging from {1.....n}and i represents the index from 1 to S. S is the size of the population (i.e, swarm). The inertia weight factor w is used in the modified PSO. The c_1 and c_2 are constants which are known as cognitive and social components, also the acceleration parameters and r_1, r_2 are the random numbers. v_{id} is the velocity vector and x_{id} is the position vector. The PSO algorithm is used in several applications such as scheduling process [9], feature selection, optimization, etc.

1.3 Fast Feature Selection Algorithm

Feature selection techniques which are previously discussed have their own advantages and disadvantages. Some have the capability to remove redundant data and some have the capability to remove irrelevant data. For example Chi square algorithm has the capability to remove irrelevant data [5]. In order to balance the both the authors have used FAST algorithm. This FAST feature selection technique has the capability to reduce both the anomalies in a balanced way.

1.4 Motivation

Feature Selection has been a key interest in recent years because of its wide uses and applications. It is pursued with great interest because of some of the following problems.

- In Datamining applications such as classification.
- Need for extraction of relevant and accurate features to analyze the huge data.
- When different parameters are merged, there is a need to integrate different such models.

2. System Architecture

2.1 Algorithm

Feature subset.

Inputs: D (F_1,F_2,....,F_m,C) the given dataset

Θ-the T-Relevance threshold.

Output: S- selected

//Irrelevant Feature removal

for i-1 to m do

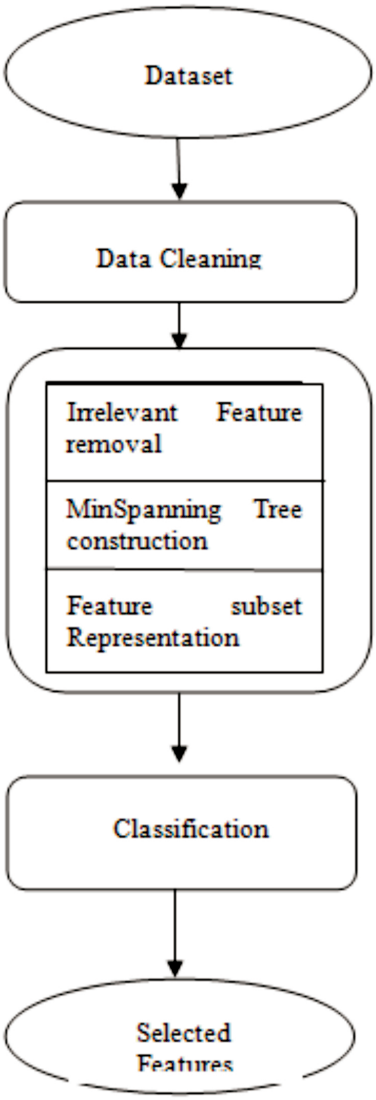

Figure 1. Data Flow Diagram of FAST Algorithm

T-Relevance = SU (F_i,C)

If T-Relevance > Θ then

S=S U({F_i};

//Minimum Spanning Tree Construction

G=NULL; //G is a complete graph

for each pair of feature {F_i',F_j'} ⊂ S do

F-Correlation- S U {F_i',F_j'}

Add F_i' and /or F_j' to G with Correlation as the weight of the corresponding edge

minSpanTree-Kruskals (G); //using Kruskal Algorithm to generate the minimum spanning tree

//Tree Partition and Feature Subset Representation

Forest=minSpanTree

for each edge E_{ij}∈Forest do

if S U (F_i',F_j')< S U (F_i,C) Ù S U (F_i',F_j')< S U (F_j',C) then

Forest= Forest- E_{ij}

for each tree T_i ∈ Forest do

F_R'=argmax F_k ∈T_i S U (F_k',C)

S= S∪ $\left(F_R^j\right)$

return S.

2.2 Description

Fast algorithm is used in order to overcome the drawback and removes the irrelevant and redundant features. The Fast algorithm works in three phases, the first phase involves irrelevant feature removal, where the irrelevant feature is removed using the T- relevance. The second phase involves construction of minimum spanning tree, this process is useful in removing the redundant data present in the dataset. The third phase involves partitioning of tree and feature selection representation.

3. Experimental Analysis

The foremost step involved in the experimental process is loading the data in the form of a dataset. Here, the authors have used the lung cancer dataset comprising of a set of attributes. During the loading process, preprocessing takes place, missing values and noisy data are removed by converting it in the ARFF (Attributed Related File Format) format.

After the conversion process, a table is created in the database, then the data is extracted from the database. Followed by extraction, the table is classified based on the last attribute. The classification is based on Naïve Bayes classification. Information gain is calculated in order to know the relevance of the attributes. Conditional entropy is calculated to know about the additional information about the attribute. The SU (Symmetric Uncertainty) is calculated by sharing the mutual gain information.

Hence entropy, conditional entropy and gain is calculated. In the next step, T-Relevance is calculated. During this process, irrelevant and redundant data are removed.

```
---------------------------------------
A15 : 1
---------------------------------------
Feature : A1------->2.545748
Feature : A2------->3.7088485
Feature : A3------->3.208323
Feature : A4------->3.7778597
Feature : A5------->3.3314996
Feature : A6------->4.074685
Feature : A7------->2.9970355
Feature : A8------->3.1701734
Feature : A9------->4.0034366
Feature : A10------->1.528525
Feature : A11------->1.9615045
Feature : A12------->0.37267798
Feature : A13------->0.4229525
Feature : A14------->3.0631769
```

Figure 2. Standard Deviation Table for Class Label 1

```
---------------------------------------
A15 : 2
---------------------------------------
Feature : A1------->3.0964808
Feature : A2------->2.2352135
Feature : A3------->3.5813897
Feature : A4------->4.429311
Feature : A5------->3.4471183
Feature : A6------->3.9624891
Feature : A7------->3.8904068
Feature : A8------->3.7975292
Feature : A9------->4.1462345
Feature : A10------->2.4793913
Feature : A11------->3.1152792
Feature : A12------->0.45907626
Feature : A13------->0.44088012
Feature : A14------->1.3520547
```

Figure 3. Standard Deviation for Class Label 2

In the next phase, f-correlation is calculated. The strong correlation between the attributes signify those features that are more relevant. Minimum spanning tree is constructed using Kruskal's algorithm and unnecessary edges are eliminated. The Kruskal's algorithm follows a greedy approach. The unnecessary edges are removed where the edge value is less than that of the threshold value. This step is very much helpful in removing redundancy.

Standard deviation are calculated in order to find the information about the neighboring feature which have the high relevance which is given in Figures 2 and 3. Information gain and conditional entropy are calculated which is used to measure the mutual information between the features (shown in Figure 4).

3.1 Redundancy Removal

Redundant data are removed after the T-relevance calculation. The following attributes subset are obtained after the relevance calculation i.e., {A2, A3, A4, A5, A6, A9, A14} are obtained for class 1 label after redundancy removal and {A3, A4, A5, A6, A7, A8, A9} for class 2 label. Based upon these obtained attribute set, the F correlation is calculated. Figure 4 represents the F-Correlation values for the attributes of class label 1 and Figure 5 represents the F-Correlation values for the class label 2. The clusters are formed based upon their correlations.

Correlations are generally used to know the linear relationship or monotonic relationship between two variables. As correlations are calculated between features, it is called F-Correlation which is shown in the Figure 5 and Figure 6. Uncertainty reduction is done through mutual information concept. Mutual information concept uses information gain concept in order to know the information content present in the concept. The main

feature	entropy	c_entropy	gain	t_relevance	class_label
A1	-118.5443478...	0.0083999809...	-118.5527478...	2.0001417187...	1
A2	-22.05215598...	0.0443029591...	-22.09645893...	2.0040180161...	1
A3	-31.72228691...	0.0310214055...	-31.75330831...	2.0019558114...	1
A4	-25.57785273...	0.0383219027...	-25.61617463...	2.0029964909...	1
A5	-28.06426071...	0.0349899928...	-28.09925070...	2.0024935624...	1
A6	-19.40114833...	0.0501915531...	-19.45133988...	2.0051740806...	1
A7	-59.48368092...	0.0166692244...	-59.50035014...	2.0005604637...	1
A8	-47.84820253...	0.0206794961...	-47.86888203...	2.0008643792...	1
A9	-21.46452431...	0.0454859990...	-21.51001031...	2.0042382489...	1
A10	-993062.2360...	0.0000010069...	-993062.2360...	2.0000000000...	1
A11	-6948.771558...	0.0001438999...	-6948.771702...	2.0000000414...	1
A12	-649.4236700...	0.0015386407...	-649.4252087...	2.0000047384...	1
A13	-439.8890246...	0.0022707147...	-439.8912954...	2.0000103240...	1
A14	-28.92483402...	0.0339677348...	-28.95880175...	2.0023486900...	1
A1	-71.34225541...	0.0276372033...	-71.36989261...	1.0003873889...	2
A2	-336.2923569...	0.0059294864...	-336.2982864...	1.0000176319...	2
A3	-50.60197938...	0.0387325677...	-50.64071195...	1.0007654358...	2
A4	-37.38317565...	0.0520426504...	-37.43521830...	1.0013921409...	2
A5	-56.29038358...	0.0348912441...	-56.32527483...	1.0006198437...	2
A6	-44.69196590...	0.0437341766...	-44.73570007...	1.0009785690...	2
A7	-49.93134099...	0.0392418705...	-49.97058287...	1.0007859166...	2
A8	-54.11456040...	0.0362670866...	-54.15082748...	1.0006701909...	2
A9	-43.93015428...	0.0444743741...	-43.97462866...	1.0010123882...	2
A10	-551.1232815...	0.0036223594...	-551.1269039...	1.0000065726...	2
A11	-107.7299311...	0.0183915377...	-107.7483226...	1.0001707189...	2
A12	-718.8209328...	0.0027784598...	-718.8237113...	1.0000038653...	2
A13	-788.5393761...	0.0025331157...	-788.5419092...	1.0000032124...	2
A14	-30531094999...	0.0000000000...	-30531094999...	1.0	2

Figure 4. Entropy, Conditional Entropy, Gain and T-Relevance Calculation

Class_Label :1

F-correlation of A2 and A3 --------->3.384716533155573
F-correlation of A2 and A4 --------->2.712314545267278
F-correlation of A2 and A5 --------->2.9695404262445666
F-correlation of A2 and A6 --------->2.174516532714867
F-correlation of A2 and A9 --------->2.336745261525383
F-correlation of A2 and A14 --------->3.0634210857361244
F-correlation of A3 and A4 --------->2.060742773039036
F-correlation of A3 and A5 --------->2.182543976116654
F-correlation of A3 and A6 --------->1.8247413592818427
F-correlation of A3 and A9 --------->1.8922243461455246
F-correlation of A3 and A5 --------->2.182543976116654
F-correlation of A3 and A6 --------->1.8247413592818427
F-correlation of A3 and A9 --------->1.8922243461455246
F-correlation of A3 and A14 --------->2.2279340779446026
F-correlation of A4 and A5 --------->2.578851485416701
F-correlation of A4 and A6 --------->1.9936388001578114
F-correlation of A4 and A9 --------->2.1101052882116202
F-correlation of A4 and A14 --------->2.649625988537017
F-correlation of A5 and A6 --------->1.9090886509866694
F-correlation of A5 and A9 --------->2.0021902573033543
F-correlation of A5 and A14 --------->2.4450684134189076
F-correlation of A6 and A9 --------->2.597882144098679
F-correlation of A6 and A14 --------->3.5241977468807613
F-correlation of A9 and A14 --------->3.1519120213655554

Figure 5. F-Correlation of Class Label 1

Class_Label :2

F-correlation of A3 and A4 --------->1.967579714975494
F-correlation of A3 and A5 --------->2.6105816575122596
F-correlation of A3 and A6 --------->2.1801142875158988
F-correlation of A3 and A7 --------->2.3614712459191276
F-correlation of A3 and A8 --------->2.521956736254624
F-correlation of A3 and A9 --------->2.1556776181311696
F-correlation of A4 and A5 --------->3.5645407865187577
F-correlation of A4 and A6 --------->2.7912828873059734
F-correlation of A4 and A7 --------->3.1223636055724935
F-correlation of A4 and A8 --------->3.408550005240367
F-correlation of A4 and A9 --------->2.7458488648519204
F-correlation of A4 and A6 --------->2.7912828873059734
F-correlation of A4 and A7 --------->3.1223636055724935
F-correlation of A4 and A8 --------->3.408550005240367
F-correlation of A4 and A9 --------->2.7458488648519204
F-correlation of A5 and A6 --------->2.0428641142670805
F-correlation of A5 and A7 --------->2.1864044536819263
F-correlation of A5 and A8 --------->2.3148935883379664
F-correlation of A5 and A9 --------->2.023700224688792
F-correlation of A6 and A7 --------->2.620724630922508
F-correlation of A6 and A8 --------->2.825914728605567
F-correlation of A6 and A9 --------->2.3544329243815754
F-correlation of A7 and A8 --------->2.551179957675578
F-correlation of A7 and A9 --------->2.1745599134350138
F-correlation of A8 and A9 --------->2.0688126432416714

Figure 6. F-Correlation of Class Label 2

intent is to select only the required variables which account for more information, i.e "80-20" rule, where 20% of variable accounts for 80% of information.

The same process is tested upon different datasets such as cancer, heart and lung cancer datasets. The obtained results are shown in the form of graphical representation for both Particle Swarm Optimization and FAST algorithms.

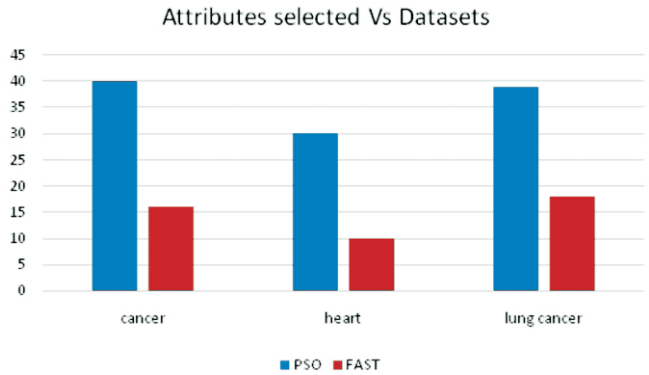

Figure 7. Comparison of Attributes Selected by PSO and FAST

The graph in the Figure 7 shows the number of attributes selected vs Datasets. The FAST algorithm produces less number of attributes which are more effective than the PSO algorithm.

Conclusion

In the present context, the authors have experimented the feature selection process using FAST feature selection algorithm. Although, there are several feature selection algorithms, the proposed feature selection algorithm has shown the positive results compared to the rest of the feature selection algorithms such PSO, Relief, Chi-square, etc. They found out that this technique more effectively selects the required effective attributes by avoiding redundant and irrelevant attributes. Hence, it increases the accuracy and reduces the preprocessing time.

References

[1]. J. Kennedy and Eberhart, (1995). "Particle Swarm Optimization". *IEEE Transactions on International Conference on Neural Networks,* Piscataway, NJ, pp. 1942-1948.

[2]. David Waha and Richard L. Bankert, (1996). "A Comparative Evaluation of Sequential Feature Selection Algorithms". *Springer-Verlag*, pp. 199-206.

[3]. Anil Jain and Douglas Zongker, (1997). "Feature Selection: Evaluation, Application and Small Sample Performance". *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol.19, No. 2, pp. 153-158.

[4]. M. Dash, and H. Liu, (1997). "Feature Selection for Classification". *Elsevier, Intelligent Data Analysis,* Vol. 1, pp. 131-156.

[5]. Huan Liu and Rudy Setiono, (1997). "Chi2: Feature Selection and Discretization of Numeric Attributes". *IEEE Transactions,* pp. 388-391.

[6]. Yuhui Shi and Russell Eberhart, (1998). "A Modified Particle Swarm Optimizer". *IEEE Transactions, Evolutionary Computation Proceedings,* pp. 69-73.

[7]. MohdSaberi Mohamad, Safaai Deris, Safie Mat Yatim, and Muhammad Razib Othman, (2004). "Feature Selection Method using Genetic Algorithm for Classification of Small and High Dimension Data". *First International Symposium on Information and Communications Technologies,* pp. 7-8.

[8]. Muhammad Imran, Rathiah Hashima and Noor Elaiza AbdKhalidb, (2013). "An Overview of Particle Swarm Optimization". *Elsevier, Procedia Engineering* , Vol. 53, pp. 491-496.

[9]. S Gracia Galan, R P Prado, Je Munoz Esposito, (2015). "Rules Discovery in Fuzzy Classifier System with PSO for Scheduling in Grid Computational Infrastructure". Elsevier, Applied Soft Computing, Vol. 29, pp. 424-435.

[10]. Simon Fong, Raymond Wong, and Athanasios V. Vasilakos, (2015). "Accelerated PSO Swarm Search Feature Selection for Data Stream Mining Big Data". *IEEE Transactions on Service Computing,* Vol. 9, No. 1, pp. 33-45.

ABOUT THE AUTHORS

K. Bhaskar Naik, is currently working as an Assistant Professor in the Department of Computer Science and Engineering, at SreeVidyanikethan Engineering College, Tirupati, India. His research area includes Image Processing, Big Data Analytics, and Datamining. He has published several papers in International and National Conferences as well as Journals in reputed Publications.

S.P. Sindhuja is currently a PG Scholar in the Department of Computer Science and Engineering at Sree Vidyanikethan Engineering College, Tirupati, India. Her research interest includes Big Data Analytics, Datamining, and Cloud Computing.