IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

Cluster Based Subset Selection Methodology Using FAST Decrees

S. Surya

Ph.D Research Scholar, J.J. College of Arts and Science, Pudukkottai, Tamil Nadu, India.

Abstract: Feature selection process plays a vital role in data mining domain, which engrosses recognizing a subset of the good number of practical features that constructs well-matched outcomes as the innovative complete deposit of features. In this paper the algorithm called Feature Selection could be experimented by means of both competence and usefulness. At the same time as the competence apprehensions the time obligatory to come across a subset of features, the usefulness is associated to the eminence of the subset of features. An innovative algorithm called a "Fast Cluster Based Feature Selection (FAST)" is proposed and the experimental results show that FAST not only produces lesser subsets of features but also get better the presentations of the classifiers.

Keywords: Feature Extraction, Subset, Clusters, FAST, Filtering Process.

I. INTRODUCTION

The main focusing of selecting a subset of high-quality features by means of administration to the objective perceptions and feature subset assortment is a successful technique for plummeting measurement, eliminating neither here nor there information, ever-increasing erudition correctness and civilizing consequence unambiguousness.

FAST algorithm employments in two phases. In the first phase, features are alienated into clusters by using graph theoretic clustering process. In the second phase, the mainly diplomat feature which is powerfully connected to objective groups is chosen from every cluster to appearance a subset of features. Features in dissimilar clusters are moderately self-governing; the clustering supported approach of FAST have an elevated likelihood of constructing a subset of practical and self-governing features. To guarantee the competence of FAST, the competent minimum spanning tree clustering Process is espoused. The competence and usefulness of the FAST algorithm are appraised from side to side an experimental revise.

II. LITERATURE REVIEW

As per the review from the paper Selective sampling approach to active feature selection [1], the Feature selection, as a preprocessing step to machine learning [1], has been very effective in reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility. We present a formalism of selective sampling based on data variance, and apply it to a widely used feature selection algorithm Relief. Further, we show how it realizes active feature selection and reduces the required number of training instances to achieve time savings without performance deterioration. We design objective evaluation measures of performance, conduct extensive experiments using both synthetic and benchmark data sets, and observe consistent and significant improvement. We suggest some further work based on our study and experiments.

As per the review from the paper Feature selection algorithms[2]: A survey and experimental evaluation, in view of the substantial number of existing feature selection algorithms, the need arises to count on criteria that enables to adequately decide which algorithm to use in certain situations. This work assesses the performance of several fundamental algorithms found in the literature in a controlled scenario. A scoring measure ranks the algorithms by taking into account the amount of relevance, irrelevance and redundancies on sample data sets [2]. This measure computes the degree of matching between the output given by the algorithm and the known optimal solution. Sample size effects are also studied.

As per the review from the paper, an introduction to variable and feature selection, Variable and feature selection [3] have become the focus of much research in areas of application for which datasets with tens or hundreds of thousands of variables are available. These areas include text processing of internet documents, gene expression array analysis, and combinatorial chemistry.

The objective of variable selection is three-fold: improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data. The contributions of this special issue cover a wide range of aspects of such problems: providing a better definition of the objective function, feature construction, feature ranking, multivariate feature selection, efficient search methods, and feature validity assessment methods.

As per the review from the paper Generalization as Search, Artificial Intelligence [4], the problem of concept learning, or forming a general description of a class of objects given a set of examples and non-examples, is viewed here as a search problem. Existing programs that generalize from examples are characterized in terms of the classes of search strategies that they employ. Several classes of search strategies are then analyzed and

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

compared in terms of their relative capabilities and computational complexities.

As per the review from the paper Feature Selection for Classification [5], *Intelligent Data Analysis*, Feature selection has been the focus of interest for quite some time and much work has been done. With the creation of huge databases and the consequent requirements for good machine learning techniques, new problems arise and novel approaches to feature selection are in demand. This survey is a comprehensive overview of many existing methods from the 1970's to the present. It identifies four steps of a typical feature selection method, and categorizes the different existing methods in terms of generation procedures and evaluation functions, and reveals hitherto un-attempted combinations of generation procedures and evaluation functions. Representative methods are chosen from each category for detailed explanation and discussion via example. Benchmark datasets with different characteristics are used for comparative study. The strengths and weaknesses of different methods are explained. Guidelines for applying feature selection methods are given based on data types and domain characteristics.

III. GRAPH THEORETICAL PROCESS

In cluster examination, graph theoretical processes have been able-bodied intentional and worn in numerous submissions. Their consequences encompass, from time to time, the most excellent conformity with person concert [17]. The universal graph theoretical clustering is straightforward: Calculate an environs graph of illustrations then remove any border in the graph that is much longer or shorter than its nationals.

Procedure: *GetUniqueItems()*

1) Create Function called "UniqueItmset()"
2) Make Exceptional Try...Catch block
3) Dimensionate the variables i, j, n and status as Integer
4) Assign the value "1" to integer declarations.
5) Dimensionate unistring as String
6) Write Clear() function to clear all the items available in the Unique Item List
7) Add the first index item into List
8) Create Looping Statements for identifying unique items

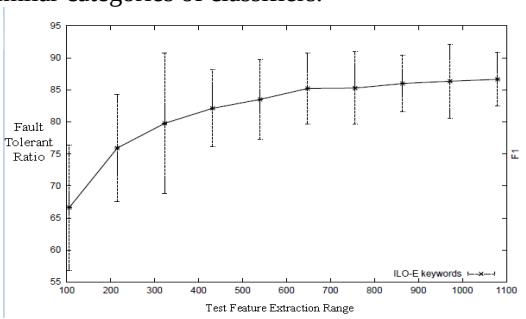
Ex:

```
for (i = 1;i<=nItems - 1;i++)
{
status = 0
for(j = 0;j<=n - 1;j++)
{
 if(UniqueItms.Item(j) == Items.Item(i))
{
 status = 1;
 }
 }
if(status == 0)
{
 UniqueItms.Add(Items.Item(i));
 n = n + 1;
 }
 }
oacalculation.setnUniqueItms(n);
for(i = 0;i<=n - 1;i++)
{
unistring = unistring + " " + UniqueItms.Item(i);
}oacalculation.setUniqueItmset(unistring);
```

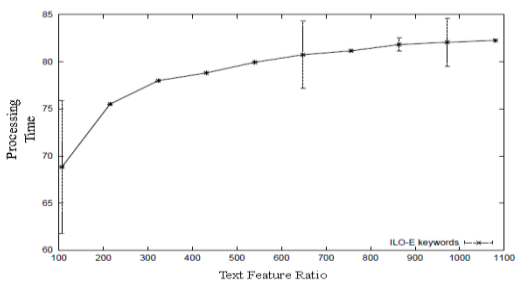
The consequence is a forest like structure and every tree in the forest symbolizes a cluster. In this learning, we pertains graph theoretical clustering processes to features. In meticulous, we assume the lowest amount spanning tree (MST) pedestaled clustering processes, because they do not take for granted that information points are collections around middles or estranged by a normal arithmetical bend and have been extensively second-hand in put into practice.

The advanced MST method is proposed, which is a Fast clustering-bAsed feature Selection algoriThm (FAST).The FAST algorithm employments in two stages. In the first stage, features are alienated into clusters by using graph theoretical clustering techniques. In the second stage, the most delegate feature that is powerfully related to objective classes is preferred from each cluster to appearance the concluding subset of features.

Features in far removed from clusters are moderately self-governing; the clustering based approach of FAST has a far above the ground likelihood of constructs a subset of practical and self-governing features. The planned feature subset assortment method FAST was experienced upon publicly available information, microarray, and text data sets. The investigational consequences demonstrate that, contrasted with additional five dissimilar categories of feature subset assortment methods, the projected method not only decreases the amount of features, but also progresses the performance of the four glowing recognized dissimilar categories of classifiers.



Graph-1: Text Categorization with Fault Tolerance Visualization



Graph-2: Text Feature Extraction Ratio with Processing Time Synchronization

DOI 10.17148/IJARCCE.2015.41234

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

IV. TEXT CLASSIFICATION BASED ON FEATURE SELECTION PROCEDURES

Document illustration and feature assortment is the majority of imperative assignment which requirements to be proficient earlier than any classification assignment. At the same time as feature assortment is also advantageous in supplementary classification assignments which is more than ever significant in text classification outstanding to the far above the ground dimensionality of text features and the continuation of neither here nor there features.

In all-purpose the text can be corresponded to in two disconnect techniques. The first is as a container of words, in which a document is corresponded to as a position of words, collectively with their connected occurrence in the document. Such a demonstration is fundamentally autonomous of the succession of words in the compilation. The second technique is corresponding to text in a straight line as strings, in which every document is a succession of words.

A good number of text categorization techniques make use of the collection of words demonstration because of its straightforwardness for categorization principles.

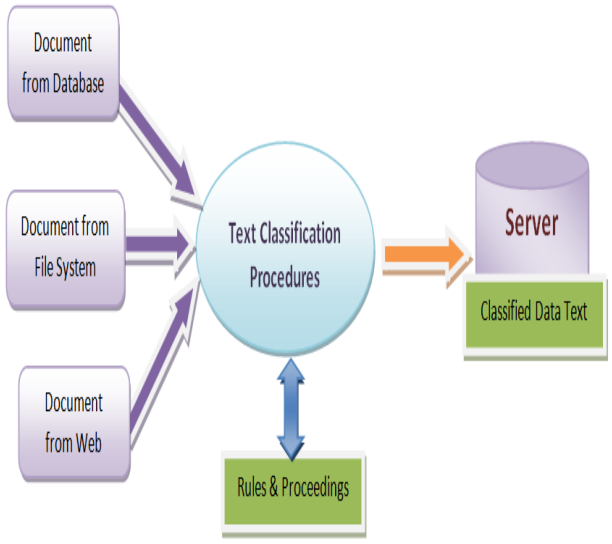


Fig.1. Text Classification Process with Documents from Various Sources

In this paper, we will talk about some of the techniques which are worn for feature assortment in text classification. The most widespread feature assortment which is used in both administered and unsubstantiated submissions is that of stop word taking away and branching. In stop word elimination, we conclude the ordinary words in the documents which are not explicit or prejudiced to the dissimilar classes. In branching process, dissimilar appearances of the identical word are combined into a solitary word. For instance remarkable, plural and diverse tenses are combined into a solitary word. We memorandum that these techniques are not exact to the case of the categorization difficulty and are frequently used in an assortment of unsubstantiated submissions such as clustering and indexing

In the case of the classification dilemma, it constructs intelligence to administer the characteristic assortment procedure with makes use of the class labels. These varieties of assortment development make sure that those features which are exceedingly twisted towards the occurrence of a meticulous class sticker are selected for the erudition progression.

Algorithm: *TextClassifyProcess()*

1) Create a Function called "getTransaction"
2) Make a function to return string values
3) Dimensionate the Integer variable "i"
4) Dimensionate the String variable called "Listitem"
5) Create a loop for calculating Number of Items.
6) Return the Gathered Items to String variable "Listitem"

Ex:

```
Function gettransaction() As String
Dim i As Integer
Dim Listitem As String = ""
For i = 0 To nitemsets - 1
Listitem = Listitem + " " + transet.Item(i)
Next
Return (Listitem)
End Function
```

Algorithm: *SubSetGeneration()*

1) Create Integer Variables "i, totuni, t"
2) Assign value "1" to all integer variable
3) Get the Unique Items by means of the function call "getnUniqueItms()"
4) Formulate a loop for constructing SubSet from the Itemset and Created Unique Items
5) Return the SubSet values to user

Ex:

```
Dim i As Integer, totuni As Integer=1
Dim t As Integer = 1
Dim merge As String = " ", disp As String = " "
totuni = getnUniqueItms() - 1
For i = 0 To totuni
subsetarray.Add(oItemset.uniarraylist(i))
farray.Add(oItemset.uniarraylist(i))
farraycount = farraycount + 1
Next
For i = 0 To totuni
        findsubarray()
Next
Return subsetarray
Sub findsubarray()
Dim i As Integer, scount As Integer, j As Integer, calculate As Integer
Dim merge As String, farray1 As New ArrayList
Dim farray2 As New ArrayList, tWords() As String
Dim k As Integer, subindex As Integer, t As Integer
For i = 0 To farraycount - 1
tWords = Split(farray.Item(i), " ")
t = UBound(tWords)
For k = 0 To getnUniqueItms() – 1
 If tWords(t) = oItemset.uniarraylist(k) Then
 subindex = k
```

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

```
 End If
 Next
scount = secondarray(subindex)
For j = 0 To scount – 1
merge = farray(i) + " " + sarray(j)
farray1.Add(merge)
subsetarray.Add(merge) calculate = calculate + 1
 Next
sarray.Clear()
 Next
 farraycount = 0
 farray.Clear()
For i = 0 To calculate - 1
farray.Add(farray1(i))
 Next
 remove(calculate)
 End Sub
```

V. EXPERIMENTAL OUTCOMES

In our experimental work, we experimentally evaluate the effectiveness of the proposed technique. The objective of our proposal is to evaluate the method in term of speed and number of selected attributes for a particular classifier on selected feature. For this application we require the following system configurations, Pentium dual core 2.4GHz system, RAM 2 GB, Hard disk of 250 GB, Monitor, and the system is implemented by means of Microsoft Visual Studio framework with the backend of SQL Server in any version, which should be greater than or equal to 2008.

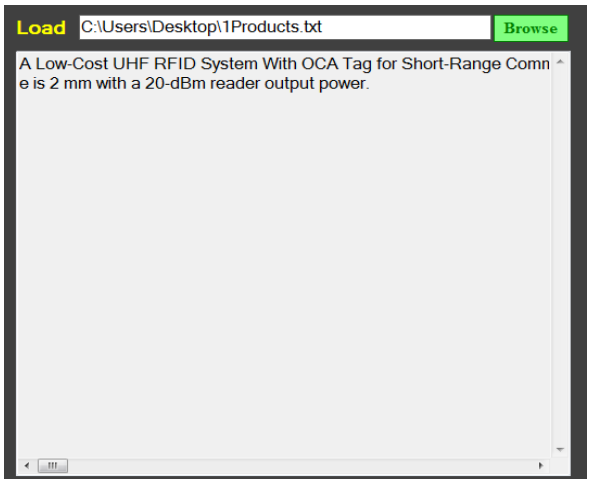


Fig 2: Open the Dataset for Processing

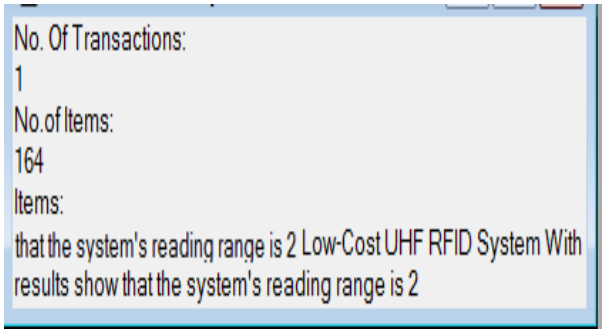


Fig 3: Calculate Transactions and Find the Items from Dataset

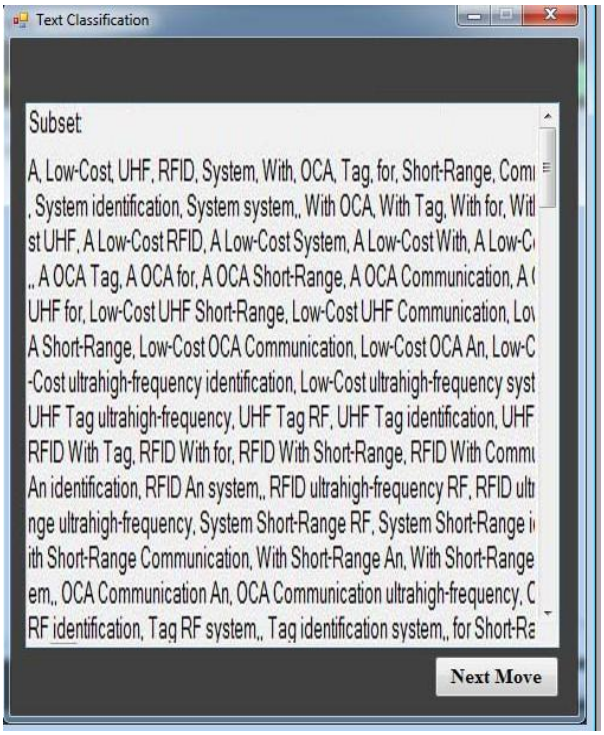


Fig 4: Subset Formulation

VI. CONCLUSION

In this system, a narrative clustering based feature subset assortment technique is proposed for elevated measurement of information. The technique engages (a) eliminate unrelated features, (b) assemble a smallest amount spanning tree beginning comparative ones and (c) separation of the subset and choosing diplomat features.

In the projected algorithm, a cluster contains of features. The classification dilemma is one of the majorities of elementary predicaments in the machine learning and data mining journalism. In the circumstance of text information, the dilemma can also be painstaking comparable to that of arrangement of disconnected set appreciated characteristics, when the frequencies of the words are unnoticed. The domains of these sets are rather outsized, as it comprises the complete glossary. Consequently, text mining procedures require to be intended to professionally administer great information of rudiments with varying frequencies. Every cluster is extravagance as a solitary feature and thus the measurement is radically abridged.

REFERENCES

[1] Liu H., Motoda H. and Yu L., Selective sampling approach to active feature selection, Artif. Intell., 159(1-2), pp 49-74 (2004).
[2] Molina L.C., Belanche L. and Nebot A., Feature selection algorithms: A survey and experimental evaluation, in Proc. IEEE Int. Conf. Data Mining, pp 306-313, 2002.
[3] Guyon I. and Elisseeff A., An introduction to variable and feature selection, Journal of Machine Learning Research, 3, pp 1157-1182, 2003.
[4] Mitchell T.M., Generalization as Search, Artificial Intelligence, 18(2), pp 203-226, 1982.
[5] Dash M. and Liu H., Feature Selection for Classification, Intelligent Data Analysis, 1(3), pp 131-156, 1997.

[6] Souza J., Feature selection with a general hybrid algorithm, Ph.D, University of Ottawa, Ottawa, Ontario, Canada, 2004.
[7] Langley P., Selection of relevant features in machine learning, In Proceedings of the AAAI Fall Symposium on Relevance, pp 1-5, 1994.
[8] Ng A.Y., On feature selection: learning with exponentially many irrelevant features as training examples, In Proceedings of the Fifteenth International Conference on Machine Learning, pp 404-412, 1998.
[9] Das S., Filters, wrappers and a boosting-based hybrid for feature Selection, In Proceedings of the Eighteenth International Conference on Machine Learning, pp 74-81, 2001.
[10] Xing E., Jordan M. and Karp R., Feature selection for high-dimensional genomic microarray data, In Proceedings of the Eighteenth International Conference on Machine Learning, pp 601-608, 2001.
[11] Souza J., Feature selection with a general hybrid algorithm, Ph.D, University of Ottawa, Ottawa, Ontario, Canada, 2004.
[12] Yu J., Abidi S.S.R. and Artes P.H., A hybrid feature selection strategy for image defining features: towards interpretation of optic nerve images, In Proceedings of 2005 International Conference on Machine Learning and Cybernetics, 8, pp 5127-5132, 2005.
[13] Dash M. and Liu H., Feature Selection for Classification, Intelligent Data Analysis, 1(3), pp 131-156, 1997.
[14] Pereira F., Tishby N. and Lee L., Distributional clustering of English words, In Proceedings of the 31st Annual Meeting on Association For Computational Linguistics, pp 183-190, 1993.
[15] Baker L.D. and McCallum A.K., Distributional clustering of words for text classification, In Proceedings of the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval, pp 96-103, 1998.
[16] Dhillon I.S., Mallela S. and Kumar R., A divisive information theoretic feature clustering algorithm for text classification, J. Mach. Learn. Res., 3, pp 1265-1287, 2003.
[17] Jaromczyk J.W. and Toussaint G.T., Relative Neighborhood Graphs and Their Relatives, In Proceedings of the IEEE, 80, pp 1502-1517, 1992.

DOI 10.17148/IJARCCE.2015.41234