ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Effective Feature Selection For High Dimensional Data using Fast Algorithm

[I]P.Abinaya, [II]Dr.J.Sutha
[I,II]Dept. of CSE, Sethu Institute of Technology, Affiliated to Anna University, Kariapatti, Tamilnadu, India

Abstract
Feature subset clustering is a powerful technique to reduce the dimensionality of feature vectors for text classification. In this paper, we propose a similarity-based self-constructing algorithm for feature clustering with the help of K-Means strategy. The words in the feature vector of a document set are grouped into clusters, based on similarity test. Words that are similar to each other are grouped into the same cluster, and make a head to each cluster data sets.
By the FAST algorithm, the derived membership functions match closely with and describe properly the real distribution of the training data. Besides, the user need not specify the number of extracted features in advance, and trial-and-error for determining the appropriate number of extracted features can then be avoided. Experimental results show that our FAST algorithm implementation can run faster and obtain better-extracted features than other methods.

Keywords
Subset, K-Means, FAST, Clustering

I. Introduction

With the aim of choosing a subset of good features with respect to the target concepts, feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility. Many feature subset selection methods have been proposed and studied for machine learning applications. They can be divided into four broad categories: the Embedded, Wrapper, Filter, and Hybrid approaches. The embedded methods incorporate feature selection as a part of the training process and are usually specific to given learning algorithms, and therefore may be more efficient than the other three categories.
Traditional machine learning algorithms like decisión trees or artificial neural networks are examples of embedded approaches. The wrapper methods use the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subsets, the accuracy of the learning algorithms is usually high. However, the generality of the selected features is limited and the computational complexity is large. The filter methods are independent of learning algorithms, with good generality.
Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed. The hybrid methods are a combination of filter and wrapper methods, by using a filter method to reduce search space that will be considered by the subsequent wrapper. They mainly focus on combining filter and wrapper methods to achieve the best possible performance with a particular learning algorithm with similar time complexity of the filter methods. The wrapper methods are computationally expensive and tend to overfit on small training sets. The filter methods, in addition to their generality, are usually a good choice when the number of features is very large. With respect to the filter feature selection methods, the application of cluster analysis has been demonstrated to be more effective than traditional feature selection algorithms. Pereira Baker and Dhillon employed the distributional clustering of words to reduce the dimensionality of text data. In cluster analysis, graph-theoretic methods have been well studied and used in many applications.
Their results have, sometimes, the best agreement with human performance. The general graph-theoretic clustering is simple: Compute a neighborhood graph of instances, then delete any edge in the graph that is much longer/shorter (according to some criterion) than its neighbors. The result is a forest and each tree in the forest represents a cluster. In our study, we apply graphtheoretic clustering methods to features. In particular, we adopt the minimum spanning tree (MST) based clustering algorithms, because they do not assume that data points are grouped around centers or separated by a regular geometric curve and have been widely used in practice. Based on the MST method, we propose a Fast clustering-bAsed feature Selection algoriThm(FAST).

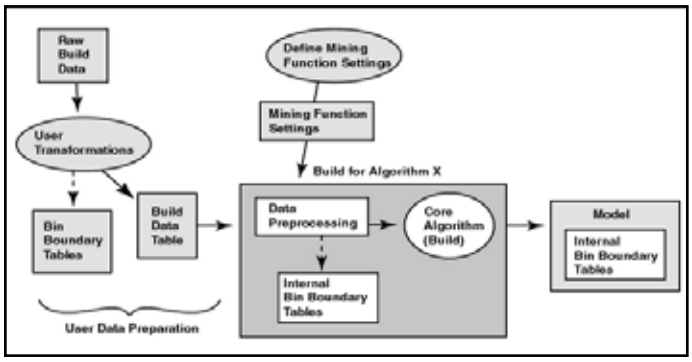


Fig. 1: MSTmethods

The FAST algorithm works in two steps. In the first step, features are divided into clusters by using graph-theoretic clustering methods. In the second step, the most representative feature that is strongly related to target clases is selected from each cluster to form the final subset of features. Features in different clusters are relatively independent; the clustering-based strategy of FAST has a high probability of producing a subset of useful and independent features. The experimental results show that, compared with other five different types of feature subset selection algorithms, the proposed algorithm not only reduces the number of features, but also improves the performances of the four well-known different types of classifiers. good feature subset is one that contains features highly correlated with the target, yet uncorrelated with each other. For all the above mentioned terms this system is a fast filter method which can identify relevant features as well as redundancy among relevant features without pairwise correlation analysis, and iteratively picks features which maximize

their mutual information with the class to predict, conditionally to the response of any feature already picked. Different from these algorithms, our proposed FAST algorithm employs clustering based method to choose features.

II. Existing System

In the past approach there are several algorithm which illustrates how to maintain the data into the database and how to retreive it faster, but the problem here is no one cares about the database maintenance with ease manner and safe methodology.

A Distortion algorithm, which creates an individual area for each and every word from the already selected transactional database, those are collectively called as dataset, which will be suitable for a set of particular words, but it will be problematic for the set of records.

A Blocking algorithm make propagation to the above problem, and reduce the problems occurred in the existing distortion algorithm, but here also having the problem called data overflow, once the user get confused then they can never get the data back.

The embedded methods incorporate feature selection as a part of the training process and are usually specific to given learning algorithms, and therefore may be more efficient than the other three categories. Traditional machine learning algorithms like decision trees or artificial neural networks are examples of embedded approaches.

The wrapper methods use the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subsets, the accuracy of the learning algorithms is usually high. However, the generality of the selected features is limited and the computational complexity is large. The filter methods are independent of learning algorithms, with good generality. Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed. The hybrid methods are a combination of filter and wrapper methods by using a filter method to reduce search space that will be considered by the subsequent wrapper. They mainly focus on combining filter and wrapper methods to achieve the best possible performance with a particular learning algorithm with similar time complexity of the filter methods.

A. Drawbacks of Existing System

- Lacks speed
- Security Issues
- Performance Related Issues
- The generality of the selected features is limited and the computational complexity is large.
- Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed.

So the focus of our new system is to enhance the throughput for any basis to eliminate the data security lacks therein and make a newer system prominent handler for handling data in an efficient manner.

III. Proposed System

Feature subset selection can be viewed as the process of identifying and removing as many irrelevant and redundant features as possible. This is because irrelevant features do not contribute to the predictive accuracy and redundant features do not redound to getting a better predictor for that they provide mostly information which is already present in other feature(s). Of the many feature subset selection algorithms, some can effectively eliminate irrelevant features but fail to handle redundant features yet some of others can eliminate the irrelevant while taking care of the redundant features.

Our proposed FAST algorithm falls into the second group. Traditionally, feature subset selection research has focused on searching for relevant features. A well-known example is Relief which weighs each feature according to its ability to discriminate instances under different targets based on distance-based criteria function. However, Relief is ineffective at removing redundant features as two predictive but highly correlated features are likely both to be highly weighted. Relief-F extends Relief, enabling this method to work with noisy and incomplete data sets and to deal with multiclass problems, but still cannot identify redundant features.

A. Advantages

1. Good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with each other.
2. The efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset.

B. In our proposed FAST algorithm, it involves

1. The construction of the minimum spanning tree (MST) from a weighted complete graph;
2. The partitioning of the MST into a forest with each tree representing a cluster;
3. The selection of representative features from the clusters.

IV. Problem Definiton

Several algorithms which illustrates how to maintain the data into the database and how to retreive it faster, but the problem here is no one cares about the database maintenance with ease manner and safe methodology. The systems like Distortion and Blocking algorithm, which creates an individual area for each and every word from the already selected transactional database, those are collectively called as dataset, which will be suitable for a set of particular words, but it will be problematic for the set of records, once the user get confused then they can never get the data back. The wrapper methods use the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subsets, the accuracy of the learning algorithms is usually high. Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed.

A FAST algorithm research has focused on searching for relevant features. A well-known example is Relief which weighs each feature according to its ability to discriminate instances under different targets based on distance-based criteria function. However, Relief is ineffective at removing redundant features as two predictive but highly correlated features are likely both to be highly weighted. Relief-F extends Relief, enabling this method to work with noisy and incomplete data sets and to deal with multiclass problems, but still cannot identify redundant features. he template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

V. Literature Review

Literature survey is the most important step in software development process. Before developing the tool it is necessary to determine the time factor, economy and company strength. Once these things are satisfied, then the next step is to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior programmers, from book or from websites. Before building the system the above consideration are taken into account for developing the proposed system.

The major part of the project development sector considers and fully survey all the required needs for developing the project. For every project Literature survey is the most important sector in software development process. Before developing the tools and the associated designing it is necessary to determine and survey the time factor, resource requirement, man power, economy, and company strength. Once these things are satisfied and fully surveyed, then the next step is to determine about the software specifications in the respective system such as what type of operating system the project would require, and what are all the necessary software are needed to proceed with the next step such as developing the tools, and the associated operations.

A. Text Classification Algorithm

The Text classification contains many problems, which has been widely studied in the data mining, machine learning, database, and information retrieval communities with applications in a number of diverse domains, such as target marketing, medical diagnosis, news group filtering, and document organization. The text classification technique assumes categorical values for the labels, though it is also possible to use continuous values as labels. The latter is referred to as the regression modeling problem. The problem of text classification is closely related to that of classification of records with set-valued features; however, this model assumes that only information about the presence or absence of words is used in a document. In reality, the frequency of words also plays a helpful role in the classification process, and the typical domain-size of text data (the entire size) is much greater than a typical set-valued classification problem.

B. Association Rule Mining

In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness. Based on the concept of strong rules, Rakesh Agrawal introduced association rules for discovering regularities between products in large-scale transaction data recorded by point-of-sale (POS) systems in supermarkets. For example, the rule {mathrm{onions, potatoes}} Rightarrow{mathrm{burger}} found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, he or she is likely to also buy hamburger meat. Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing or product placements. In addition to the above example from market basket analysis association rules are employed today in many application areas including Web usage mining, intrusion detection, Continuous production and bioinformatics. As opposed to sequence mining, association rule learning typically does not consider the order of items either within a transaction or across transactions.

C. FAST Algorithm

FAST Algorithm is a classic algorithm for frequent item set mining and association rule learning over transactional databases. This FAST algorithm inbuiltly contains an algorithm called Apriori, which proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis.

D. Feature Selection Algorithm

In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in model construction. The central assumption when using a feature selection technique is that the data contains many redundant or irrelevant features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context.

Feature selection techniques are a subset of the more general field of feature extraction. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the features. Feature selection techniques are often used in domains where there are many features and comparatively few samples (or datapoints).

Feature selection techniques provide three main benefits when constructing predictive models:

- Improved model interpretability,
- Shorter training times,
- Enhanced generalisation by reducing overfitting.

Feature selection is also useful as part of the data analysis process, as shows which features are important for prediction, and how these features are related.

VI. System Architecture

A. User Module

In this module, Users are having authentication and security to access the detail which is presented in the ontology system. Before accessing or searching the details user should have the account in that otherwise they should register first.

B. Distributed Clustering

The Distributional clustering has been used to cluster words into groups based either on their participation in particular grammatical relations with other words by Pereira et al. or on the distribution of class labels associated with each word by Baker and McCallum . As distributional clustering of words are agglomerative in nature, and result in suboptimal word clusters and high computational cost, proposed a new information-theoretic divisive algorithm for word clustering and applied it to text classification. Proposed to cluster features using a special metric of distance, and then makes use of the of the resulting cluster hierarchy to choose the most relevant attributes. Unfortunately, the cluster evaluation measure based on distance does not identify a feature subset that allows the classifiers to improve their original performance accuracy. Furthermore, even compared with other feature selection methods, the obtained accuracy is lower.

C. Subset Selection Algorithm

The Irrelevant features, along with redundant features, severely affect the accuracy of the learning machines. Thus, feature subset selection should be able to identify and remove as much of the irrelevant and redundant information as possible. Moreover, "good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other. Keeping these in mind, we develop a novel algorithm which can efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset.

D. Text Classification Process

The Text classification contains many problems, which has been widely studied in the data mining, machine learning, database, and information retrieval communities with applications in a number of diverse domains, such as target marketing, medical diagnosis, news group filtering, and document organization. The text classification technique assumes categorical values for the labels, though it is also possible to use continuous values as labels. The latter is referred

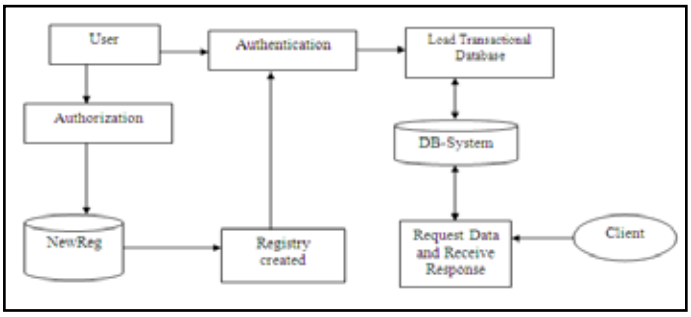


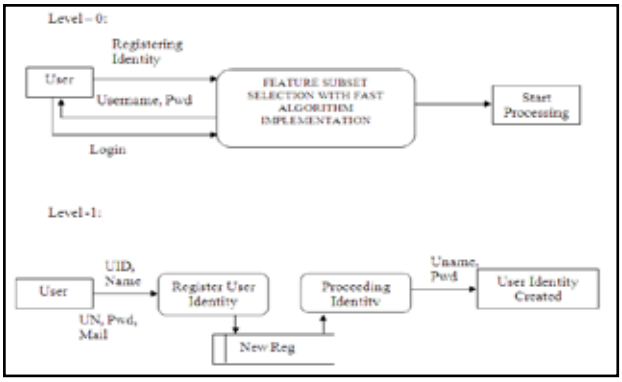
Fig. 2.Text Classification Process

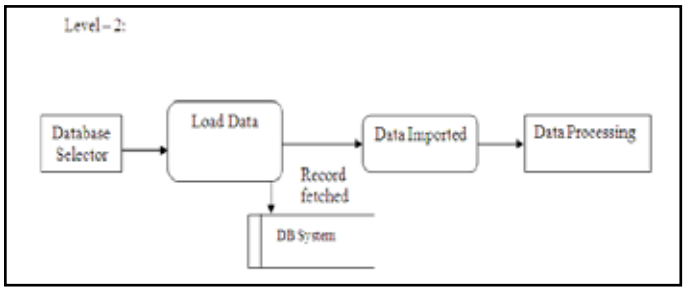
Classification is closely related to that of classification of records with set-valued features; however, this model assumes that only information about the presence or absence of words is used in a document. In reality, the frequency of words also plays a helpful role in the classification process, and the typical domain-size of text data (the entire size) is much greater than a typical set-valued classification problem.

E. Feature Selection Algorithm

Feature selection techniques are a subset of the more general field of feature extraction. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the features. Feature selection techniques are often used in domains where there are many features and comparatively few samples (or datapoints).

VII. Data Flow Diagram





VIII. Sample Codings

A. Load Data

```
Dim tResult As Integer
        OpenFileDialog1.Filter = "Text files (*.txt)|*.txt|" + "All
files|*.*"
        OpenFileDialog1.FileName = Application.StartupPath() +
"\data\*.txt"
      tResult = OpenFileDialog1.ShowDialog()
      If tResult = Windows.Forms.DialogResult.OK Then
         CurrentDBPath = OpenFileDialog1.FileName
         txtpath.Text = CurrentDBPath
        Dim tExtension As String = LCase(Mid(CurrentDBPath,
InStr(CurrentDBPath, ".") + 1))
         If tExtension = "txt" Then
            LoadDB_txt()
Cont..
```

B. Text Classification

```
Sub UniqueItmset()
         Try

      tResult = OpenFileDialog1.ShowDialog()
      If tResult = Windows.Forms.DialogResult.OK Then
         CurrentDBPath = OpenFileDialog1.FileName
         txtpath.Text = CurrentDBPath
         Dim tExtension As String = LCase(Mid(CurrentDBPath
  If tExtension = "txt" Then
            LoadDB_txt()
Cont..
```

C. Required Classes

```
      Sub UniqueItmset()
         Try
            Dim i As Integer, j As Integer, n As Integer = 1, status
As Integer
            Dim unistring As String = ""
            UniqueItms.Clear()
            UniqueItms.Add(Items.Item(0))
            For i = 1 To nItems - 1
               status = 0
               For j = 0 To n - 1
                  If UniqueItms.Item(j) = Items.Item(i) Then
                     status = 1
                  End If
               Next
               If status = 0 Then
                  UniqueItms.Add(Items.Item(i))
                  n = n + 1
               End If
            Next
```

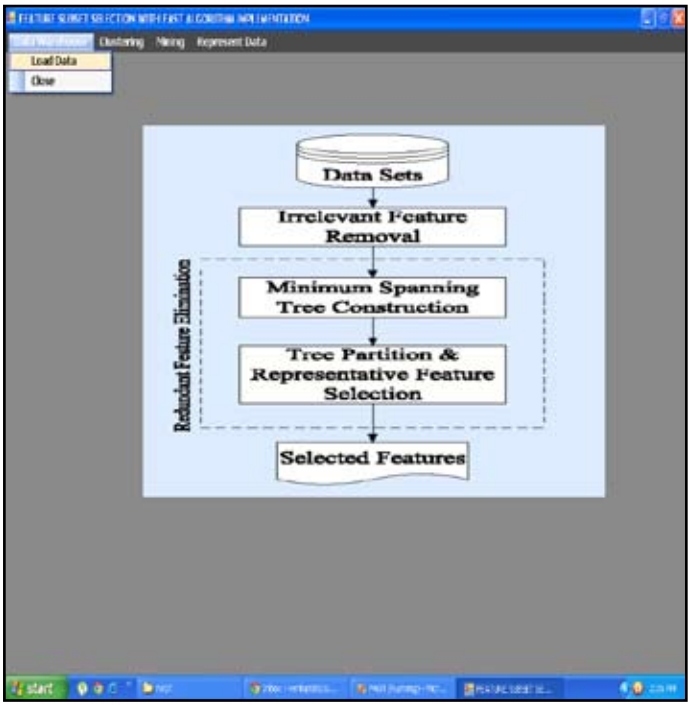
ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

```
oacalculation.setnUniqueItms(n)
For i = 0 To n - 1
   unistring = unistring + " " + UniqueItms.Item(i)
Next
oacalculation.setUniqueItmset(unistring)
```

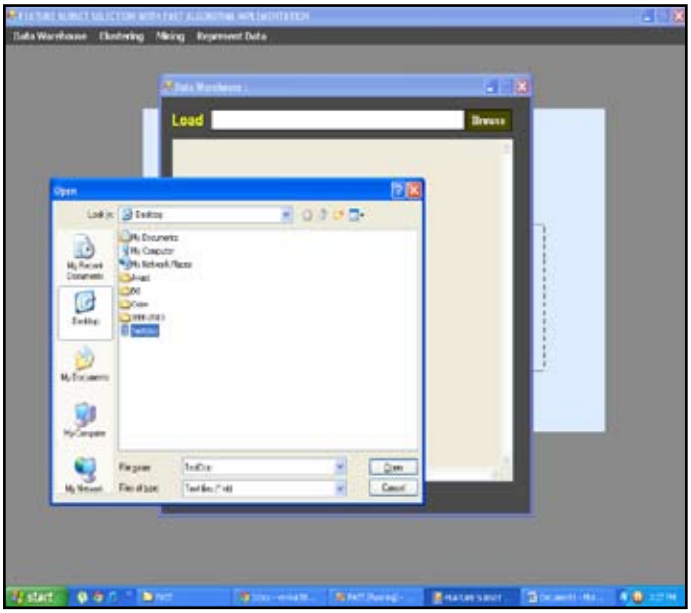
Cont.

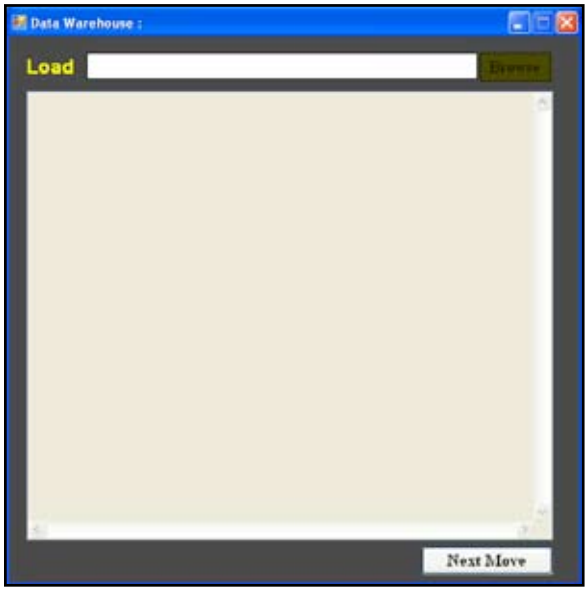
IX. Screen Shots

A.Main Form

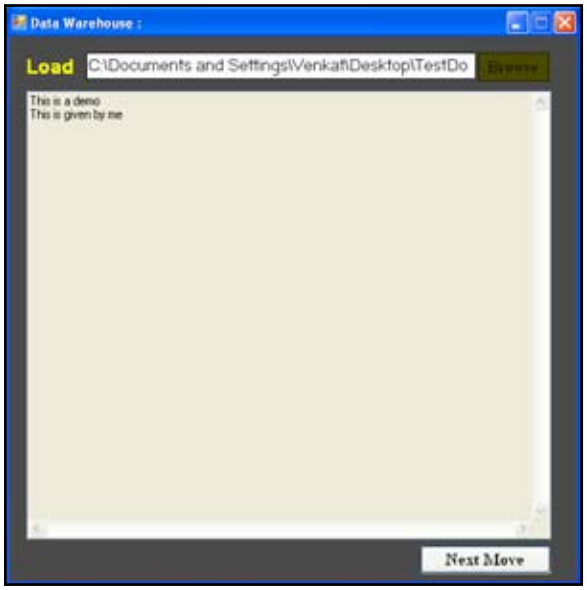


B.Load Transactional Data





C. Select the data



D. Text Classification

Acknowledgment

The research work has defined a new rule set namely the informative rule set that presents prediction sequences equal to those presented by the association rule set using the confidence priority. The informative rule set is significantly smaller than the association rule set, especially when the minimum support is small. The proposed work has characterized the relationships between the informative rule set and the non-redundant association rule set, and revealed that the informative rule set is a subset of the non-redundant association rule set. The work considers the upward closure properties of informative rule set for omission of uninformative association rules, and presented a direct algorithm to efficiently generate the informative rule set without generating all frequent item sets.

The informative rule set generated in our work is significantly smaller than both the association rule set and the non-redundant association rule set for a given database that can be generated more efficiently than the association rule set. The efficiency improvement results from that the generation of the informative rule set needs fewer candidates and database accesses than that of the association rule set rather than large memory usage like some other algorithms. The number of database accesses of the proposed algorithm is significantly fewer than other direct methods for generating association rules on all items. So far we have identified that the performance of quantitative algorithms is considerably lower than the proposed methodologies.

References

[1] Almuallim H. and Dietterich T.G., Learning boolean concepts in the presence of many irrelevant features, Artificial Intelligence, 69(1-2), pp 279-305, 1994.

[2] 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.

[3] Dash M. and Liu H., Feature Selection for Classification, Intelligent Data Analysis, 1(3), pp 131-156, 1997.

[4] Liu H., Motoda H. and Yu L., Selective sampling approach to active feature selection, Artif. Intell., 159(1-2), pp 49-74 (2004)

[5] Park H. and Kwon H., Extended Relief Algorithms in Instance-BasedFeature Filtering, In Proceedings of the Sixth International Conference on Advanced Language Processing and Web Information Technology (ALPIT 2007), pp 123-128, 2007.

[6] W. Duch, Filter Methods. In: Feature extraction, foundations and applications. Eds: I.Guyon, S. Gunn, M. Nikravesh, L. Zadeh, Studies in Fuzziness and Soft Computing,Physica-Verlag, Springer, pp. 89-118, 2006.

[7] Raman B. and Ioerger T.R., Instance-Based Filter for Feature Selection,

Journal of Machine Learning Research, 1, pp 1-23, 2002.

[8] Souza J., Feature selection with a general hybrid algorithm, Ph.D, University of Ottawa, Ottawa, Ontario, Canada, 2004.

[9] 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.

Abinaya.P received her B.E degree in Computer Science And Engineering from Mount Zion College Of Engineering And Technology, India, in 2012. She is currently pursuing her M.E in Computer science and Engineering from Sethu Institute of Technology, India. Her research interest includes Datamining.