International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 14s,2025
https://theaspd.com/index.php

Weight Based Feature Subse t Selection from High Dimensional Databases

Dr. P.NagaKavitha[1], Shaziya Fatima[2], Barad EstherRani[3]
[1]Associate Professor, Department of Computer Applications(MCA), St.Ann's College for Women, Mehdipatnam, Osmania University, Hyderabad, Telangana, India, kavithareddy.pasam@gmail.com
[2]Assistant Professor, Department of Commerce, St.Ann's College for Women, Mehdipatnam, Osmania University, Hyderabad, Telangana, India, s_fatima2002@yahoo.com
[3]Assistant Professor, Department of Computer Science(PG- M.Sc.(AI & ML), St.Ann's College for Women, Mehdipatnam, Osmania University, Hyderabad, Telangana, India, estherbarad@gmail.com

Abstract:
The problem of curse of dimensionality is understood in the wake of data mining algorithms taking long time and consuming more resources in processing databases that contain high- dimensions. When there are more dimensions in the dataset, naturally it consumes more resources. However, there might be some attributes that are not needed for processing. In other words, it is possible to reduce dimensions by identifying representative features that provide full coverage of all attributes. In this paper we proposed an algorithm that based on weight, entropy and gain for identifying representative features in a high-dimensional data. Entropy and gain are computed systematically in order to have a weight associated with all attributes. Different pairs of attributes are correlated and representative features are selected. The representative features are the features that represent all features in the given dataset. We built a prototype application that demonstrates proof of the concept. The empirical results revealed that the proposed algorithm is effective and can be used in the data mining operations for improving performance. This is achieved by reducing dimensions through the proposed feature subset selection.
Keywords–*Datamining,high-dimensionaldata,clustering,featuresubsetselection.*

I.INTRODUCTION

Data mining is widely used in real world applications. It is the discipline where historical data is analyzed to obtain hidden information. In other words, it is the process of extracting or discovering latent trends or patterns that are not known earlier. These trends or patterns uncovered from the databases are used to take expert decisions. The process of mining is essential for any enterprise in different domains. Knowledge discovery helps domain experts to have interpretation of knowledge and take decisions. Models are built in order to have solutions to different problems. The general steps involved in knowledge discovery from databases (KDD) are is visualized in Figure 1.

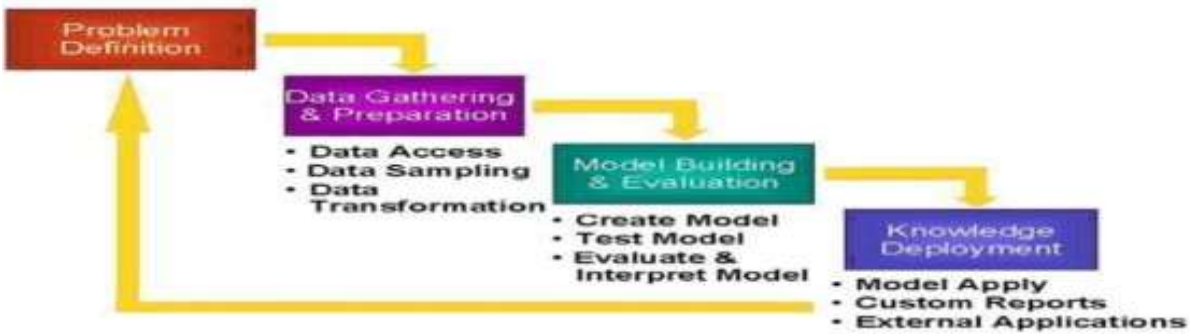


Figure1:Stepsin data mining

There are many steps in KDD. First of all a problem is defined. Then data is gathered in order to solve the problem. Then data mining algorithms are used to built a model and evaluate it. This gives rise to knowledge

needed. This knowledge is used to make expert decisions that result in business growth and profits. There are many algorithms related to data mining. They include association rule mining, decision trees, clustering and classification. These algorithms take time and resources to complete mining process. When high dimensional data is taken, these algorithms take long time to execute and consume more resources. To overcome this problem, it is important to reduce dimensions. In this paper we proposed an algorithm to extract feature subset from high dimensional data. The algorithm is meant for feature subset selection from high dimensional data. It takes given dataset T-Relevance threshold (limit), and target concept C as inputs and produces feature subsets. The functionality of the algorithm is divided into three parts. They are removal of irrelevant features, construction of minimum spanning tree (MST), and tree partition for selecting feature subsets that are representative in nature. The remainder of the paper is structured as follows. Section II provides review of literature. Section III presents the proposed system in detail. Section IV presents implementation details. Section V shows experimental results while section VI concludes the paper.

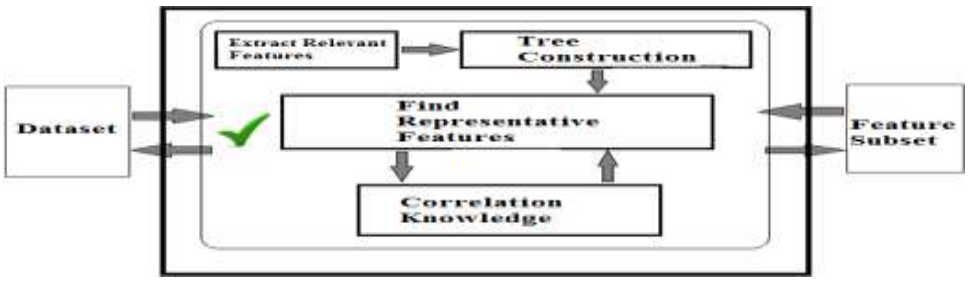
Relatedworks

This section provides review of literature on feature subset selection and other related topics. Almuallim and Dietterich [1] proposed algorithms for identification of features that are relevant to a given concept. Arauzo-Azofra et al. [2] proposed a measure known as feature set measurefor improved performance of data mining algorithms. Battiti[3]explored the process of selecting features from database using supervised learning approaches. The problems of high dimensional data are explored in[4],[5],[6], and[7]. Feature selection algorithms are studied in [8] for simplifying data mining operations. Correlation based feature subset selection is presented in [9] where machine learning concepts are used to achieve it. Similar kind of work is done in[10]. In[11] features election problems are explained with traditional methods and a novel method that differs from traditional methods. In [12], the researchers focused on wrappers for feature subject selection as part of their data mining framework. Machine learning is used in [13] for feature subset selection. In [14], information- theoretic approaches are employed for feature subset selection while in [15] probabilistic approaches are studied for the same. In [16] a sampling approach is followed for feature selection. All these approaches are used to improve data mining algorithms by reducing dimensionality. The work in this paper is also in the similar lines with highly useful representative features selected. Gain and entropy are used in order to find whether an attribute is useful for data mining purposes. Representative features are identified based on the waits given to attributes in order to reduce number of dimensions in the given high-dimensional dataset.

II. Proposedsystem

In this paper we proposed a framework and proposed an algorithm. The framework is as showninFigure2.Thedatasetistakenasinput and generates feature subset that provides representative features. This way it is made possible to reduce number of attributes. The feature subset is the representative of all features. With correlation knowledge, the representative features are used to generate final feature subset. This kind of feature subset selection which is based on clustering has its utility in the real world.

Figure 2



As shown in Figure2, the proposed system takes dataset as input and extracts relevant features. Afterwards a tree is constructed in order to process the rest of the procedure as discussed with the proposed algorithm.

The tree is used in the process
of finding representative features. Correlation knowledge is utilized while making final decisions on feature subject which is the outcome of the algorithm.

III. Proposedalgorithm

Algorithm: Weight, Entropy and Gain Based Feature Subset Algorithm
Inputs: Dataset **D**
Outputs:
Feature subset
Step1:Initialization
Initialize gain threshold, entropy threshold, feature subset vector, feature vector, concept and tree Based on given concept get attributes of D into attributes vector.
Step2:ExtractRelevantFeatures
Compute entropy and gain for all attributes Compute average weight and associate it with attributes Based on the gain and entropy thresholds add attributes to feature set.
Step3:ConstructionofTree
Add all features to a tree
Step4:FindRepresentativeFeatures
For each node in the tree Find whether any feature pair is correlated.
Step5:Output
Output feature subset
First of all initialization of required variables such as Gain threshold (limit), Entropy threshold, Feature subset vector (array), Feature vector, concept and Attributes vector are done. It gets all attributes (column headings) of dataset into an array/vector. For all attributes entropy and gain are computed. Then weight is associated with attributes. Finally attributes with given thresholds are chosen into feature vector. A tree is constructed to have all chosen features in the feature vector in an organized fashion. The tree makes traversing easier. From all the features stored in F, features are studied to find out representative features so that it is possible to eliminate some of the features which have a representative feature. This is to reduce search space and improve efficiency of clustering process in the real world. Especially it is used when data has multiple dimensions. When dimensions are reduced, processing becomes easier and computational cost is reduced. Then it returns the final feature subset which has representative features for all dimensions of high dimensional data.

V.IMPLEMENTATION AND RESULTS

We built a prototype application to demonstrate the utility of the proposed application. The application is built using Java platform. Java SWING is used to develop graphical user interface. The implementation is done using the proposed algorithm implementation. The java.io package is used to have file IO operations while java.util package is used to have collections needed. JTable is the UI control used to present the data in the application. Usability is given importance while designing the application. It is user friendly and helps users to have data mining operations like clustering for feature subset selection process.

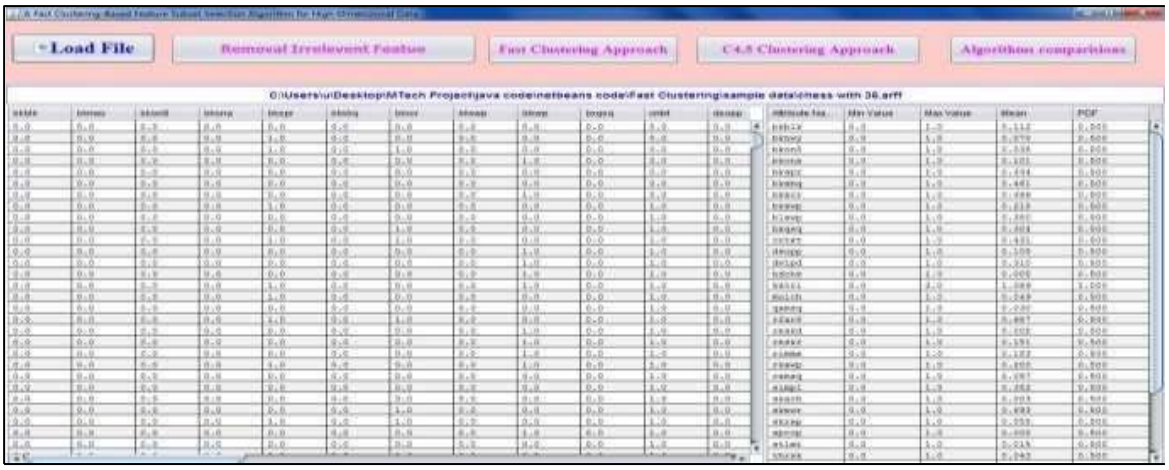


Figure3:MainUIfordifferentoperations

The main UI shown in Figure 1 provides needed graphical interface to end user. It has provision for loading data, removal of irrelevant features fast clustering of data for feature selection, C4.5 clustering approach and comparison of proposed algorithm and that of C4.5.

Figure 4: Computations of entropy and gain Entropy and gain are the two statistics or measures to identify importance of an attribute for performing data mining operations like clustering. Entropy refers to disorder or uncertainty. Entropy and Gain are the two statistical measures used to make decision pertaining to removal of irrelevant features and choosing relevant features that are relevant to the chosen concept. These two are also used to know correlation between two attributes in given dataset. Entropy characterizes impurity of an attribute while the Gain is the expected reduction in entropy when examples are partitioned according to given attribute.

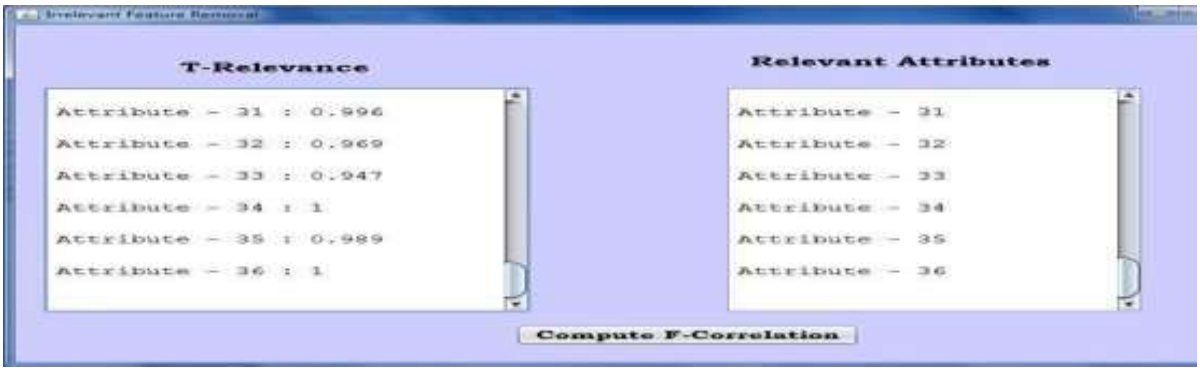


Figure5:ComputesT-Relevanceandfindrelevantattributes.

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 14s,2025
https://theaspd.com/index.php

Figure5:ComputesT-Relevanceandfindrelevantattributes.

The algorithm used in this paper is meant for feature subset selection from high dimensional data. It takes given dataset T-Relevance threshold (limit), and target concept C as inputs and produces feature subsets.

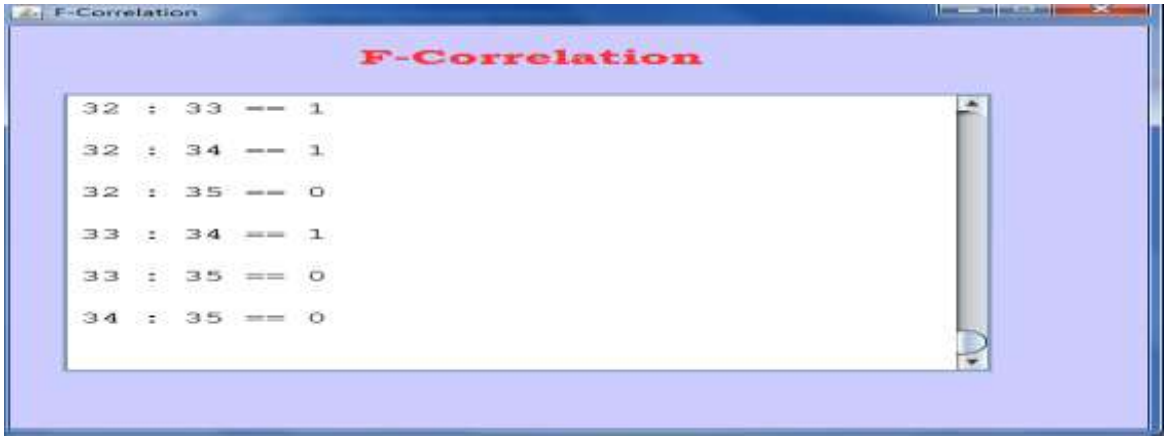


Figure6:ComputationofF-Correlation

The relevant features identified in the first part of the algorithm are used here. For every pair of features F-correlation is computed. F-Correlation is the measure to know how two features are related. The graph is constructed to represent all features with link between then and F- correlation value on the edges. F-correlation results are shown in Figure 7.

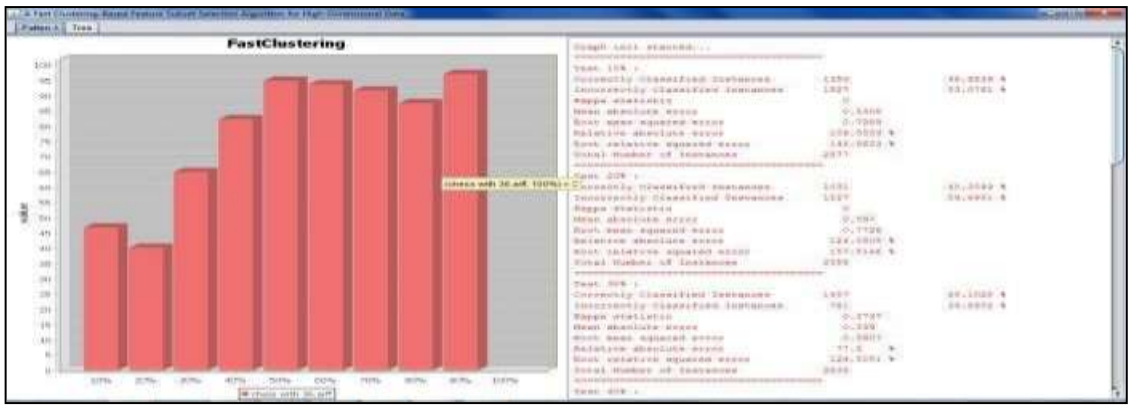


Figure7:shows proposed clustering results of proposed algorithm with different % of training data

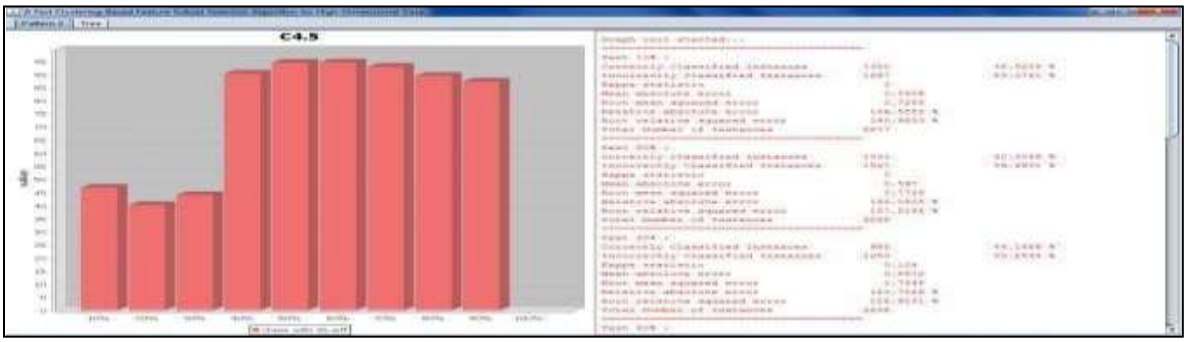


Figure8-clusteringresultsofC4.5withdifferent % of training data

As shown in Figure 8, it is evident that the clustering is made with the proposed algorithm for representative feature subset selection. It shows the test data percentage, number of instances that are classified correctly, error rate in classification and so on. The trends in the result revealed that the percentage of training data has its influence on the clustering results.

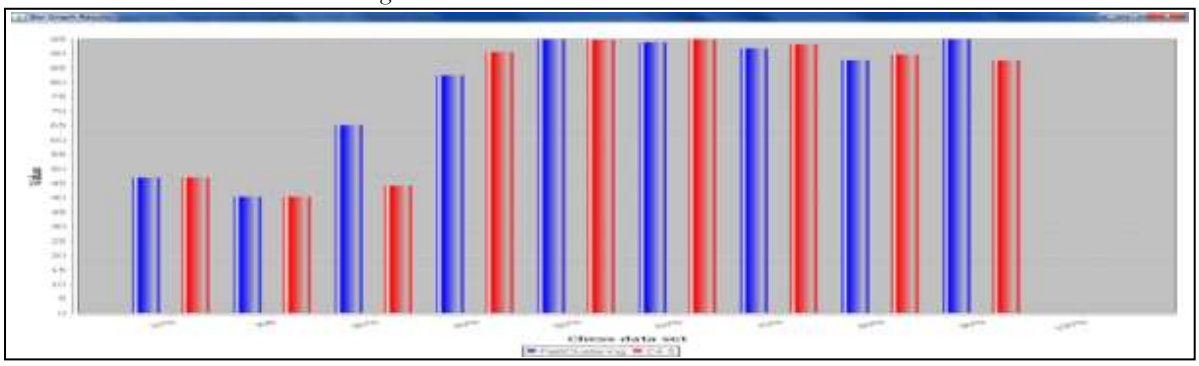


Figure9-Comparison between clustering performance of proposed algorithm and C4.5

As shown in Figure 9, it is evident that the clustering is made with the proposed algorithm for representative feature subset selection. It shows the test data percentage, number of instances that are classified correctly, error rate in classification and so on. The trends in the result revealed that the percentage of training data has its influence on the clustering results. As shown in Figure 9, it is evident that the proposed algorithm has comparable performance improvement over its predecessors that C4.5 algorithm. The results are provided with different training datasets. The results reveal that the clustering is made with different percentage of training data.

VI.CONCLUSIONS ANDFUTURE WORK

In this paper, we proposed an algorithm that based on weight, entropy and gain for identifying representative features in a high-dimensional data. Entropy and gain are computed systematically in order to have a weight associated with all attributes. Different pairs of attributes are correlated and representative features are selected. The representative features are the features that represent all features in the given dataset. Chess dataset is taken from UCI machine learning repository in order to make experiments. Webuilta prototype application to demonstrate the proof of the concept. The empirical results revealed that the proposed algorithm is effective in reporting representative features. The results obtained from the application are compared with the result of other clustering algorithm known as C4.5. There is comparable performance improvement of C4.5. In future we intend to improve the proposed algorithm for working with big data containing high dimensions.

REFERENCES

[1] H. Almuallim and T.G. Dietterich, “Algorithms for Identifying Relevant Features,” Proc. Ninth Canadian Conf. Artificial Intelligence, pp. 38-45, 1992.

[2] A.Arauzo-Azofra,J.M.Benitez,and
J.L. Castro, “A Feature Set Measure Based on Relief,” Proc. Fifth Int’l Conf. Recent Advances in Soft Computing, pp. 104-109, 2004.

[3] R. Battiti, “Using Mutual Information for Selecting Features in Supervised Neural Net Learning,” IEEE Trans. Neural Networks,vol.5,no.4,pp.537-550,July 1994.

[4] J.BiesiadaandW.Duch,“FeaturesElectionforHigh-DimensionaldataaPearso Redundancy Based Filter, ”Advances in Soft Computing,vol.45,pp.242-249,2008.

[5] E. Xing, M. Jordan, and R. Karp, “Feature Selection for High Dimensional Genomic Microarray Data,” Proc. 18th Int’l Conf. Machine Learning, pp. 601-608, 2001.

[6] L. Yu and H. Liu, “Feature Selection or High-Dimensional Data: A Fast Correlation- Based Filter Solution,” Proc. 20th Int’l Conf. Machine Leaning, vol. 20, no. 2, pp. 856-863, 2003.

[7] L. Yu and H. Liu, “Efficiently Handling Feature Redundancy in High-Dimensional Data,” Proc. Ninth ACM SIGKDD

Int'lConf. Knowledge Discovery and Data Mining (KDD '03), pp. 685-690, 2003.

[8] L.C. Molina, L. Belanche, and A. Nebot, "Feature Selection Algorithms: A Survey and Experimental Evaluation," Proc. IEEE Int'l Conf. Data Mining, pp. 306-313, 2002

[9] M.A. Hall, "Correlation-Based Feature Subset Selection for Machine Learning," PhD dissertation, Univ. of Waikato, 1999.

[10] M.A. Hall, "Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning," Proc. 17thInt'lConf.MachineLearning,pp.359- 366, 2000.

[11] K. Kira and L.A. Rendell, "TheFeature Selection Problem: Traditional Methods and a New Algorithm," Proc.10th Nat'l Conf. Artificial Intelligence, pp. 129- 134, 1992.

[12] R.KohaviandG.H.John,"WrappersforFeatureSubsetSelection,"Artificial Intelligence,vol.97,nos.1/2,pp.273-324, 1997.

[13] P. Langley, "Selection of Relevant Features in Machine Learning," Proc.AAAI Fall Symp. Relevance, pp. 1-5,1994.

[14] M.Last,A.Kandel,andO.Maimon,"Information-TheoreticAlgorithmforFeature Selection," Pattern Recognition Letters, vol. 22, nos. 6/7, pp. 799-811, 2001.

[15] H.LiuandR.Setiono,"AProbabilisticApproachtoFeatureSelection:AFilterSolution,"Proc. 13th Int'l Conf. Machine Learning, pp. 319-327, 1996.

[16] H. Liu, H. Motoda, and L. Yu, "Selective Sampling Approach to Active Feature Selection," Artificial Intelligence, vol. 159, nos. 1/2, pp. 49-74, 2004.