ISSN 23499842(Online), Volume 1, Special Issue 1(NCFCSIT 15), March 2015
International Journal of Innovative Trends and Emerging Technologies

A NOVEL FEATURE SELECTION ALGORITHM FOR HIGH DIMENSIONAL DATA USING KFGM CLUSTERING

Dr.S.ANITHAA[1] J.PRIYA[2]
[1]professor, [2]M.E Second Year, [1,2]Department of CSE
[1,2]Sri Ramanujar Engineering College

Abstract: Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus the most important information in their data warehouses. Feature selection is the process of selecting a subset of relevant features for use in model construction. The central when using a feature selection technique is that data contains many redundant or irrelevant features. Ofthe many feature subset selection algorithms, some can effectively eliminate irrelevant features but fail to handleredundant features yet some of others can eliminate their relevant while taking care of the redundant features. Our proposed FAST algorithm falls into the second group. Traditionally, feature subset selection research has focused searching for relevant features. This paper provides the implementation of this algorithm on high dimensional data.

Keyword: Data mining ,Feature Selection ,FAST Algorithm,K-Means,FCM algorithm.

1.INTRODUCTION

In Machine learning, the problem of supervised classification is concerned with using labeled examples to induce a model that classifies objects into a finite set of known classes.
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

2 .PROBLEM STATEMENT:

Feature subset selection can be viewed as the process of identifying and removing as many irrelevant and redundant features as possible. This is because 1) irrelevant features do not contribute to the predictive accuracy and 2) 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. In existing system, clustering-based feature subset selection algorithm for high dimensional data is presented. Here cluster consists of features. Each cluster is treated as a single feature and thus dimensionality is drastically reduced. But, it fails when clustering more features. Here limited number of features is used to prepare feature subset.

OBJECTIVE:

Feature selection is one of the most used techniques to reduce dimensionality among practitioners. It aims to choose a small subset of the relevant features from the original ones. The aim of this approach is to reduce computing time to detect attacks in a network. Each and every layer in the Layered architecture will be trained separately and then implemented sequentially.

3.RELATED WORK

Siddheshwar V. Patil1, Prakash J. Kulkarniproposed that intrusion detection faces a number of challenges; an intrusion detection system must reliably detect malicious activities in a network and must perform efficiently to cope with the large amount of network traffic[1].proposed that the automated process of going through large amounts of data with the intention to discover useful information about the data that is not obvious[2].Useful information may include special relations between the data, specific models that of the data that repeats itself, specific patterns, and ways of classifying it or discovering specific values that fall out of the "normal" pattern or model[3]. **Tran, D et al.,** proposed that the computing networks had become an absolute tool for various sectors which includes

social, economies, military and so on. It ensures the connectivity, collaboration and cooperation between these different sectors[4]. FenyeBaoet proposed that a highly scalable cluster-based hierarchical trust management protocol for wireless sensor networks (WSNs) to effectively deal with selfish or malicious nodes. [5]. Gerhard M¨unz, Sa Li, Georg Carle proposed that data mining techniques make it possible to search large amounts of data for characteristic rules and patterns. This paper gives an introduction to Network Data Mining, i.e. the application of data mining methods to packet and data captured in a network, including a comparative overview of existing approaches[6].Despite of growing information technology widely, security has remained one challenging area for computers and networks. technologies applied to intrusion detection to invent invent a new pattern from the massive network data as well as to reduce the strain of the manual compilations of the intrusion and normal behavior patterns[7].

4.EXISTING SYSTEM

Given a connected, undirected graph, a spanning tree of that graph is a sub graph that is a tree and connects all the vertices together. A single graph can have many different spanning trees. We can also assign a weight to each edge, which is a number representing how unfavorable it is, and use this to assign a weight to a spanning tree by computing the sum of the weights of the edges in that spanning tree. A minimum spanning tree (MST) or minimum weight spanning tree is then a spanning tree with weight less than or equal to the weight of every other spanning tree. More generally, any undirected graph (not necessarily connected) has a minimum spanning forest. A minimum spanning tree would be one with the lowest total cost. Another approach for attack based dataset feature subset selections are Decision tree method, select the best features for each decision node during the construction of the tree based on well defined criteria. This method has very high speed of operation and high attack detection accuracy.

Disadvantages

- Low accuracy in attack detection.
- High computing time.
- Require more memory.
- Computationally expensive.

5.PROPOSED SYSTEM ALGORITHMS AND ADVANTAGES

IRRELEVANT REDUNDANT FERTURE REMOVAL

The irrelevant feature removal is straightforward once the right relevance measure is defined or selected. The former obtains features relevant to the target concept by eliminating irrelevant ones. Feature subset selection can be the process that identifies and retains the strong T-Relevance features. Irrelevant features have no/weak correlation with target concept. Relevant features have strong correlation with target concept so are always necessary for a best subset, while redundant features are not because their values are completely correlated with each other. Thus, notions of feature redundancy and feature relevance are normally in terms of feature correlation and feature-target concept correlation.

REDUNDANT FEATURE

The former obtains features relevant to the target concept by eliminating irrelevant ones, and the latter removes redundant features from relevant ones via choosing representatives from different feature clusters, and thus produces the final subset.

The irrelevant feature removal is straightforward once the right relevance measure is defined or selected, while the redundant feature elimination is a bit of sophisticated, and because our proposed feature subset selection framework involves irrelevant feature removal and redundant feature elimination, we firstly present the traditional definitions of relevant and redundant features, then provide our definitions based on variable correlation as follows.

MEMBERSHIP MATRIX

For detection of attacks based feature subset selection, Adaptive fuzzy C-means anomaly detection method is proposed. For FCM clustering, the degree of membership for a pattern in a particular cluster is 1 or 0 if it doesn't . The membership grades determine the degree to which the pattern belongs to these clusters.

The degree of membership for a pattern in a particular cluster is 1 or 0 if it doesn't .In fuzzy set theory, a pattern can belong to 2 or more clusters simultaneously. The membership grades determine the degree to which the pattern belongs to these cluster.The proposed FAST algorithm logically consists of three steps: 1) removing irrelevant features, 2) constructing an MST from relative ones, and 3) partitioning the MST and selecting representative features.

If no data elements are exchanged between clusters, the process will be halted.

ADAPTIVE FUZZY C-MEANS CLUSTERING ALGORITHM

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. It is frequently used in pattern recognition. Straightly speaking, this algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster and the data point. More the data is near to the cluster center more is its membership towards the particular cluster center. Clearly, summation of membership of each data point should be equal to one.

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2 , \quad 1 \le m < \infty$$

where *m* (the Fuzziness Exponent) is any real number greater than 1, *N* is the number of data, *C* is the number of clusters, u_{ij} is the degree of membership of x_i in the cluster *j*, x_i is the *i*th of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center.

K-MEANS CLUSTERING ALGORITHM

K-means clustering is a data mining/machine learning algorithm used to cluster observations into groups of related observations without any prior knowledge of those relationships. The k-means algorithm is one of the simplest clustering techniques and it is commonly used in medical imaging, biometrics and related fields.

The advantages are its simplicity and low computational cost, which allows it to run efficiently on large datasets. The main drawback is, it won't produce the same result the algorithm runs each time and the resulting clusters depend on the initial assignments.

MINIMUM SPANNING TREE CONSTRUCTION

Algorithm steps:

1. A spanning tree of a graph is a tree and is a sub graph that contains all the vertices
2. A graph may have many spanning trees, for example the complete graph on four vertices has sixteen spanning trees.

Suppose that the edges of the graph have weight or lengths, the weight of a tree will be the sum of weight of its edges.

GRAPH THEORETIC CLUSTERING

The steps were shown below

1. Compute a neighbor graph of instance.
2. Delete the edge which is too long and too short from the graph.
3. Result is a forest where each tree in forest represent a cluster.

6.SYSTEM ARCHITECTURE DIAGRAM

The system architecture design include all the components of the fast clustering algorithm in addition to this it also include the two clustering algorithms like FCM and k-means clustering

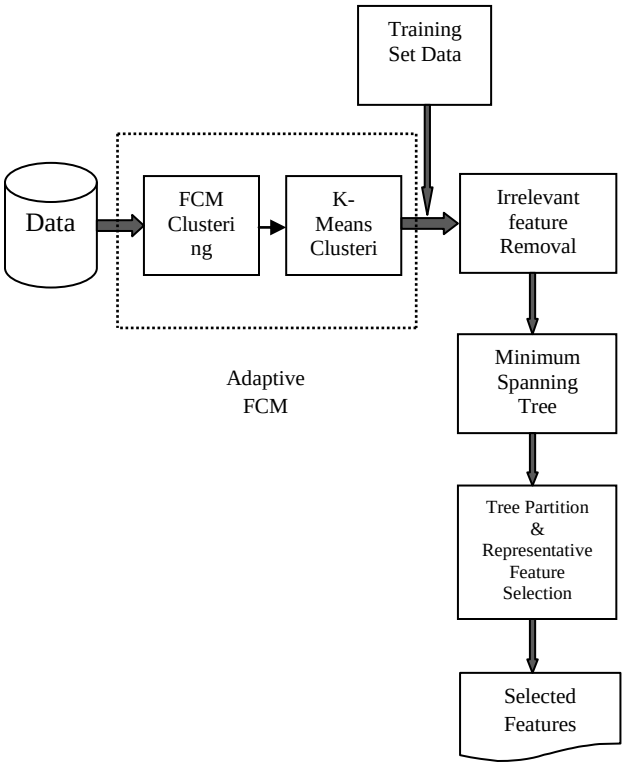
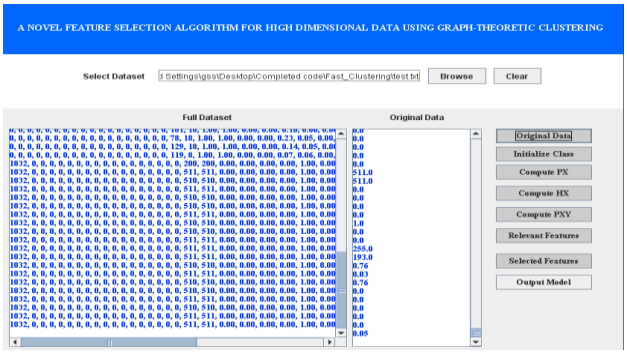


Fig1 System Architecture Design

7.IMPLEMENTATION RESULT

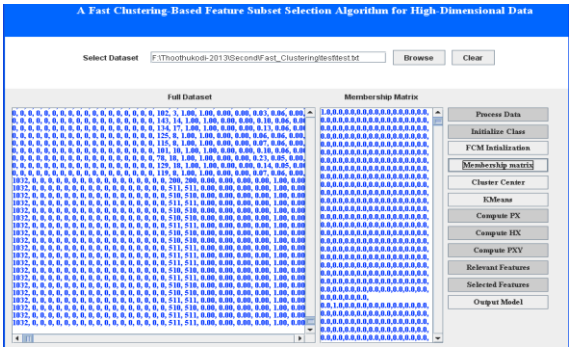
ORIGINAL DATA



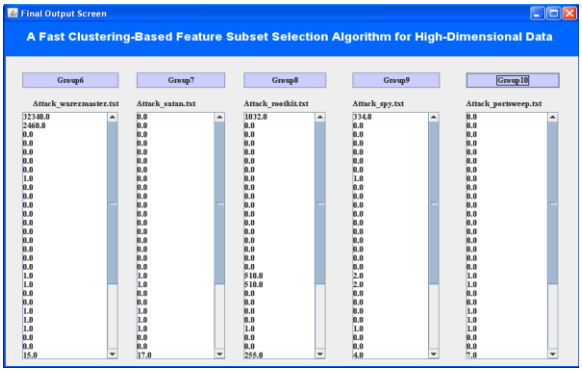
The symmetric uncertainty (SU) is derived from the mutual information by normalizing it to the entropies of feature values or feature values and target classes, and has been used to evaluate the goodness of features.

MEMBERSHIP MATRIX

Use the FCM for fuzzy partition, so that each given data instance can determine which categories belong to, according to the membership between 0 and 1. The elements of the matrix U get values between 0 and 1



FINAL OUTPUT



8.PERFORMANCE EVALUATION

TIME COMPLEXITY

Information gain is a symmetrical measure. That is the amount of information gained about X after observing Y is equal to the amount of information gained about Y after observing X. This ensures that the order of two variables will not affect the value of the measure.

Symmetric uncertainty treats a pair of variables symmetrically; it compensates for information gain's bias toward variables with more values and normalizes its value to the range [0, 1]. A value 1 of SU(X, Y) indicates that knowledge of the value of either one completely predicts the value of the other and the value 0 reveal that X and Y are independent.

The symmetric uncertainty is defined as follows:

$$SU(X,Y) = \frac{2 \times Gain(X|Y)}{H(X) + H(Y)}.$$

8.CONCLUSION AND FUTURE WORK

Feature subset selection for detecting various attacks is proposed. For detection of attacks, we have used Adaptive fuzzy C-means detection method and KDD CUP 1999 data set. We trained with 10 types of attack models with 37 labels of dataset. FAST clustering method does not capture if we give more attacks. But in Adaptive FCM method, it detects more attacks and a novel clustering-based feature subset selection algorithm for high dimensional dataThe advantages are its simplicity and low computational cost, which allows it to run efficiently on large datasets. For future work, we plan to explore different types of correlation measures and study some formal properties of feature space.

REFERENCES

1. **Almuallim.HandDietterich.T.G**, "Learning Boolean Concepts in the Presence of Many Irrelevant Features," Artificial Intelligence, vol. 69, nos. 1/2, pp. 279-305, 1994.
2. **Arauzo-Azofra.A, Benitez.J.M, and Castro.J.L**, "A Feature Set Measure Based on Relief," Proc. Fifth Int'l Conf. Recent Advances in Soft Computing, pp. 104-109, 2004.
3. **Baker.L.DandMcCallum.A.K**, "Distributional Clustering of Words for Text Classification," Proc. 21st Ann. Int'l ACM SIGIR Conf. Research and Development in information Retrieval, pp. 96-103, 1998.

4. **Battiti.R**, "Using Mutual Information for Selecting Features in Supervised Neural Net Learning," IEEE Trans. Neural Networks, vol. 5, no. 4, pp. 537-550, July 1994.
5. **Bell.D.A and Wang.H**, "A Formalism for Relevance and Its Application in Feature Subset Selection," Machine Learning, vol. 41, no. 2, pp. 175-195, 2000.
6. **Biesiada.J and Duch.W**, "Features Election for High-Dimensional data a Pearson Redundancy Based Filter," Advances in Soft Computing, vol. 45, pp. 242-249, 2008.
7. **Butterworth.R, Piatetsky.G-Shapiro, and Simovici.D.A**, "On Feature Selection through Clustering," Proc. IEEE Fifth Int'l Conf. Data Mining, pp. 581-584, 2005.
8. Cardie.C, "Using Decision Trees to Improve Case-Based Learning," Proc. 10th Int'l Conf. Machine Learning, pp. 25-32, 1993.
9. **Cohen.W**, "Fast Effective Rule Induction," Proc. 12th Int'l Conf. Machine Learning (ICML '95), pp. 115-123, 1995.
10. **Dash.MandLiu.H**, "Feature Selection for Classification," Intelligent Data Analysis, vol. 1, no. 3, pp. 131-156, 1997.
11. **Cohen.W**, "Fast Effective Rule Induction," Proc. 12th Int'l Conf. Machine Learning (ICML '95), pp. 115-123, 1995.
12. **Dash.M, Liu.H, and Motoda.H**, "Consistency Based Feature Selection," Proc. Fourth Pacific Asia Conf. Knowledge Discovery and Data Mining, pp.98-109, 2000.
13. **Das.S**, "Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection," Proc. 18th Int'l Conf. Machine Learning, pp.74- 81, 2001.
14. **Dhillon.I.S, Mallela.S**, and **Kumar.R**, "A Divisive Information TheoreticFeature Clustering Algorithm for Text Classification," J. Machine Learning Research, vol. 3, pp. 1265-1287, 2003.