A Survey on Clustering Based Feature Selection Technique Algorithm for High Dimensional Data

Mr Avinash Godase [1#1], Mrs Poonam Gupta [2. #2]
PG Student [1],

Guide[2]G. H. Raisoni college of Engineering &Management,Wagholi Pune

ABSTRACT

A high dimensional data is enormous issue in data mining and machine learning applications. Feature selection is the mode of recognize the good number of features that produce well-suited outcome as the unique entire set of features. Feature selection process constructs a pathway to reduce the dimensionality and time complexity and also improve the accuracy level of classifier Feature selection involves identifying a subset of the most useful features that produces compatible results as the original entire set of features. In the primary step, features are separated into clusters hence we are study various technique and algorithm such as of graph-theoretic clustering methods a Density-based Clustering Algorithms Distance-based Clustering ,Distributed Clustering Techniques

Keywords:-FAST, Tree based Clustering etc.

1. INTRODUCTION

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Data mining tasks are specified by its functionalities that tasks are classified into two forms:

1. Descriptive mining tasks: Portray the general properties of the data.
2. Predictive mining tasks: Perform the implication on the current data order to craft prediction

In machine learning and statistics, feature selection also known as variable selection, attribute selection or variable subset selection. It 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[01]. 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 data points. A feature selection algorithm can be seen as the combination of a search technique for proposing new feature subsets, along with an evaluation measure which scores the different feature subsets. The simplest algorithm is to test each possible subset of features finding the one which minimizes the error rate. Feature subset selection is an effectual way for dimensionality reduction, elimination of inappropriate data, rising learning accurateness, and recovering result unambiguousness. Numerous feature subset selection methods have been planned and considered for machine learning applications [2].They can be separated into four major categories such as: the Wrapper, Embedded, and Filter and Hybrid methods. In particular, we accept the minimum spanning tree based clustering algorithms, for the reason that they do not imagine that data points are clustered around centers or separated by means of a normal geometric curve and have been extensively used in tradition [3].The removal of Irrelevant features which have weak correlation and redundant features are the main key factors in feature subset selection process and effective way of feature subset selection.

2. LITERATURE SURVEY

In [1], Qinbao Song et al, proposed a new FAST algorithm that gain more accuracy and reduce time complexity than traditional feature selection algorithm like, FCBF, Relief, CFS, FOCUS-SF, Consist and also compare the classification accuracy with prominent classifiers. Graph theoretic clustering and MST based approach is used for ensure the efficiency of feature selection process. Classifiers plays vital roles in feature selection operation since accuracy of selected features are measured using the progression of classifiers. The following classifiers are utilized to classify the data sets [2], [3], Naïve Bayes: it works under Bayes theory and is based on probabilistic approach and yet then offers first-rate classification output. C4.5 is the successor of ID3 [4] support of decision tree induction method. Gain ratio, gini index information gain are the measures used for the process of attribute selection. Simplest algorithm is IB1 (instance based) [5]. Based on the distance vectors, it performs the classification process. RIPPER [6] is the rule based technique, it make a set of rules for the purpose of classify the data sets. Classifier is one of the evaluation parameter for measuring the accuracy of the process.

3.RELATED WORK

Efficient Feature Selection via Analysis of Relevance and Redundancy

This paper[4] study a new framework of feature selection which avoids implicitly handling feature redundancy and turns to efficient elimination of redundant features via explicitly handling feature redundancy. Relevance definitions divide features into strongly relevant, weakly relevant, and irrelevant ones; redundancy definition further divides weakly relevant features into redundant and non redundant ones. The goal of this paper is to efficiently find the optimal subset. We can achieve this goal through a new framework of feature selection (figure 1) composed of two steps: first, relevance analysis determines the subset of relevant features by removing irrelevant ones, and second, redundancy analysis determines and eliminates redundant features from relevant ones and thus produces the final subset. Its advantage over the traditional framework of subset evaluation lies in that by decoupling relevance and redundancy analysis, it circumvents subset search and allows a both efficient and effective way in finding a subset that approximates an optimal subset. The disadvantage of this technique is that it does not process the image data[9].

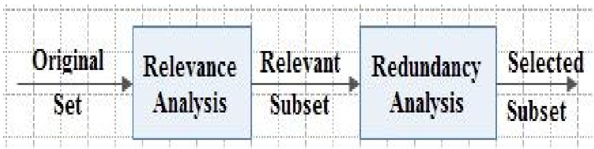


Figure 1: A new framework of feature selection

Filter approach

Filter approach uses intrinsic properties of data for feature selection. This is the unsupervised feature selection approach. This approach performs the feature selection without using induction algorithms]. This Filter method is generally used for the transformation of variable space. This transformation of variable space is required for the collation and computation of all the features before dimension reduction can be achieved[7].

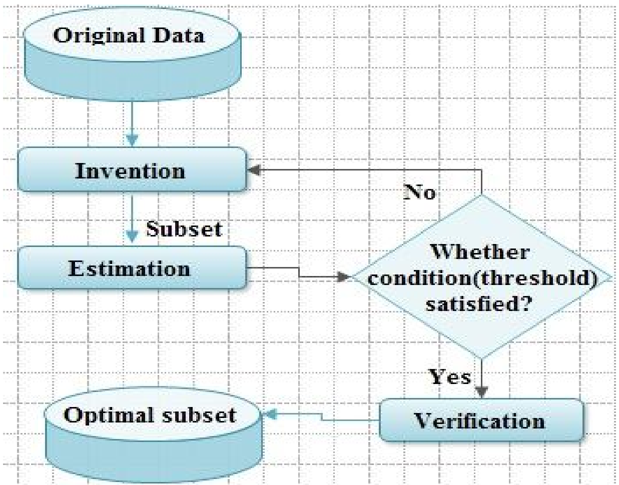


Fig 2: Feature selection process

Steps for feature selection process:

Invention*:* process with original dataset, produce a candidate subset.

Estimation*:* evaluation is done using candidate subset.

Decision process*:* Compare the selected subset with checking criteria (threshold value)

Verification*:* cross validation is performed for select optimal subset of features.

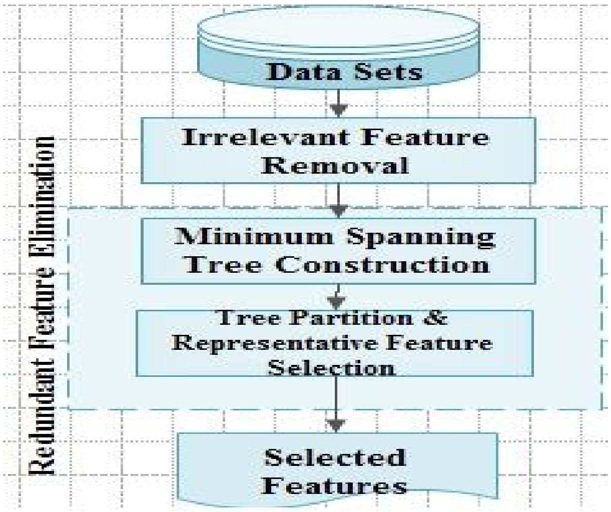

Data Sets
Irrelevant Feature Removal
Minimum Spanning Tree Construction
Tree Partition & Representative Feature Selection
Selected Features
Redundant Feature Elimination

Fig3 : Framework of the feature subset selection algorithm

A novel algorithm which can efficiently and effectively deal with both immaterial and unnecessary features, and obtain a good feature subset. We achieve this through a new feature selection framework (shown in Fig.1) which collected of the two connected components of immaterial feature removal and unnecessary feature elimination. The former obtains features applicable to the target concept by eliminating immaterial ones, and the latter removes unnecessary features from applicable ones via choosing representatives from different feature clusters, and thus produces the final subset. The immaterial feature removal is straightforward once the right relevance measure is the unnecessary feature elimination is a bit of complicated[4].Another important task is removal of unnecessary features which also affect on performance of algorithm.

Classification methods:

Bayesian Classification*:*

Bayesian classifiers are statistical classifiers used to predict class membership probabilities. It is also known as naïve Bayesian classifier based on Bayes theorem. Compare to other classifiers it have the minimum error rate[1].

Decision tree Induction*:*

Decision trees are constructed in a top-down recursive divide-and-conquer method. It consists of three algorithms such as ID3 (Iterative Dichotomiser), C4.5 (successor of ID3), CART (Classification and Regression Trees). The procedure employs an attribute selection measure such as gini index, information gain and gain ratio. Attribute selection measure [1] is used to separates the original data set (D) into individual classes[05][06].

Rule Based Classification*:*

A rule-based classifiers uses a set of rules for classification task. This method effectively produces the subset of features using different heuristic techniques[6].

Association Rule Mining:

Association rule mining is the best method for discovering interesting relations between variables in large databases or data warehouse. It is intended to identify strong rules discovered in databases using different measures of interestingness. With the help of this rule mining the system can manipulate and associates the text cluster into the respective heads based on the internal features of data[6].

Clustering

Clustering is a semi-supervised learning problem, which tries to group a set of points into clusters such that points in the same cluster are more similar to each other than points in different clusters, under a particular similarity matrix. 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).

Graph based clustering

The general methodology of graph-based clustering includes the below given five part :

(1) Hypothesis. The hypothesis can be made so that a graph can be partitioned into densely connected sub graphs that are sparsely connected to each other.

(2) Modeling. It deals with the problem of transforming data into a graph or modeling the real application as a graph by specially designating the meaning of each and every vertex, edge as well as the edge weights.

(3) Measure. A quality measure is an objective function that rates the quality of a clustering. The quality measure will identify the cluster that satisfy all the desirable properties.. An algorithm is to exactly or approximately optimize the quality measure. The algorithm can be either top down or bottom up.

(5) Evaluation. Various metrics can be used to evaluate the performance of clustering by comparing with a ―ground truth clustering[01].

Density-based Clustering Algorithms

Density-based algorithms assign each sample to all the clusters with different probabilities. Then, the clusters are defined as the areas of higher density, meaning that some samples, called noisy or border points, could remain “outside” the clusters, namely in the areas that separate the groups. The main idea of this clustering policy is to enlarge the existing cluster as long as the density in the neighborhood withdraws a certain threshold. For this reason, density based clustering algorithms are suitable in finding non-linear shaped clusters[14].

Distance-based Clustering Techniques

Distance-based algorithms analyze the dissimilarity between samples by means of a distance metric and assess the most representative pattern of each cluster, called centroid. Afterwards, the class is decided by assigning the sample to the closest centroids are found targeting small dissimilarity distances to the samples of the own cluster and large dissimilarity distances to the samples of the other clusters. Obviously, there are situations when it becomes unclear how to assign a distance measure to a data set and how to associate the weights of the features[14].

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[10]. 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.

Embedded Approaches

Embedded approaches [10], sometimes also referred to as nested subset methods [11], act as an integral part of the machine learning algorithm itself. During the operation of the classification process, the algorithm itself decides which attributes to use and which to ignore. Just like wrapper methods, embedded approaches thus depend on a specific learning algorithm, but may be more efficient in several aspects. Since they avoid retraining of the classifier from scratch for every feature subset investigated, they are usually faster than wrapper methods. Moreover, they make better use of the available data by not needing to split the training data into a training and test/validation set. Decision trees are famous examples that use embedded feature selection by selecting the attribute that achieves the “best” split in terms of class distribution at each leave. This procedure is repeated recursively on the feature subsets until some stopping criterion is satisfied.

Wrappers

Filters, wrappers [6] do use the learning algorithm as an integral part of the selection process. The selection of features should consider the characteristics of the classifier. Then, in order to evaluate subsets, wrappers use the classifier error rate induced by the learning algorithms as its evaluation function[7].

This aspect of wrappers results in higher accuracy performance for subset selection than simple filters. Wrappers have to train a classifier for each subset evaluation, they are often much more time consuming. The main type of evaluation methods are

i. Distance (Euclidean distance measure).
ii. Information (entropy, information gain, etc.)
iii. Dependency (correlation coefficient).
iv. Consistency (min-features bias).
v. Classifier error rate (the classifier themselves).

After a feature subset is generated, it is feed into an evaluation process where the process will compute some kind of relevancy value. The generated subset candidate is feed into an evaluation function it will compute some relevancy value The generation steps are able to categories different feature selection method according to the way evaluation is carried out. The first four consider as a filter approach and the final one as a wrapper approach.

Relief Algorithm

Relief is well known and good feature set estimator. Feature set estimators evaluate features individually. The fundamental idea of Relief algorithm [4], [5] is estimate the quality of subset of features by comparing the nearest features with the selected features. With nearest hit (H) from the same class and nearest miss (M) from the different class perform the evaluation function to estimate the quality of features. This method used to guiding the search process as well as selecting the most appropriate feature set. Relief estimates are better than usual statistical attribute estimates, like correlation or covariance because it consider attribute interrelationships. It is better to use a reduced set of features.

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[04].

Flow chart:

The following Diagram shows the flow chart for implementing the clustering based feature selection algorithm.

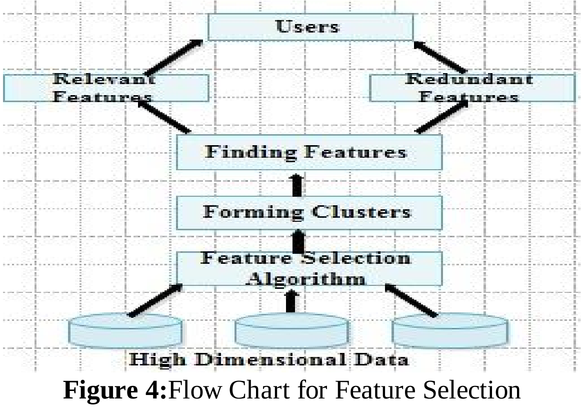


Figure 4: Flow Chart for Feature Selection

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 features which provide

No more information than the currently selected features, and irrelevant features provide no useful information in any context.Feature selection techniques provide three main benefits when constructing predictive models:

- Improved model interpretability,
- Shorter training times,
- Enhanced generalization 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.

MST Construction

With the F-Correlation value computed above, the Minimum Spanning tree is constructed. For that, we use Kruskal□s algorithm which form MST effectively. Kruskal's algorithm is a greedy algorithm in graph theory that finds a minimum spanning tree for a connected weighted graph. This means it finds a subset of the edges that forms a tree that includes every vertex, where the total weight of all the edges in the tree is minimized. If the graph is not connected, then it finds a minimum spanning forest (a minimum spanning tree for each connected component)[1].

FAST Algorithm

FAST is Tree-Based Algorithm and Advanced Chameleon is Graph-Based Algorithm. Features in different clusters are relatively independent; the clustering-based strategy of Chameleon has a high probability of producing a subset of useful and independent features. To ensure the efficiency of FAST, we adopt the efficient minimum-spanning tree clustering method, for Chameleon we adopt the K – Nearest neighbor graph clustering method. The efficiency and effectiveness of the FAST and Chameleon algorithms are evaluated through an empirical study. Feature subset selection algorithms, most of them can effectively eliminate irrelevant features but fail to handle redundant features. The FAST method undergoes the Feature selection process by undergoing four phases[01][04].

A. Irrelevant Feature Removal

This phase is concerned with the removal of irrelevant features that does not match with the target concept. The features are extracted as irrelevant using a match concept that reveals the relevance property between a feature and its target class. If there is no match between the values of the selected feature f and the target class c, it is said to be irrelevant and thus removed from the set of features. If the relevance measure is beyond the threshold then that feature is selected[11].

B. Clustering

The Distributional clustering has been used to cluster words into groups based either on their participation in particular grammatical relations with other words or on the distribution of class labels associated with each word. As distributional clustering of words is agglomerative in nature, and result is suboptimal word clusters and acquires high computational cost[02][03].

C. Redundant Feature Removal

The next phase in FAST method is redundant feature removal. After removing the irrelevant features, there occurs need to remove the redundant features. If a feature is embedded with redundant information, then it may not contribute to the better prediction of target classes. Redundant features completely correlate with each other. So if F is a set of features then it is said to be redundant if it has Markov Blanket within F. Assuming this as the redundant feature is removed. The major amount of work for FAST algorithm involves the computation of Symmetric Uncertainty (SU) values from which the T-Relevance that is relevance between feature and target concept and F-Correlation that is relevance between any pair of features are calculated. This measure has linear complexity in terms of the number of instances in a given data set. The first part of the algorithm has a linear time complexity in terms of the number of features m. The Improved FAST algorithm strives to improve the time complexity by reducing the time taken to calculate the SU values thus increasing the overall performance[01][04].

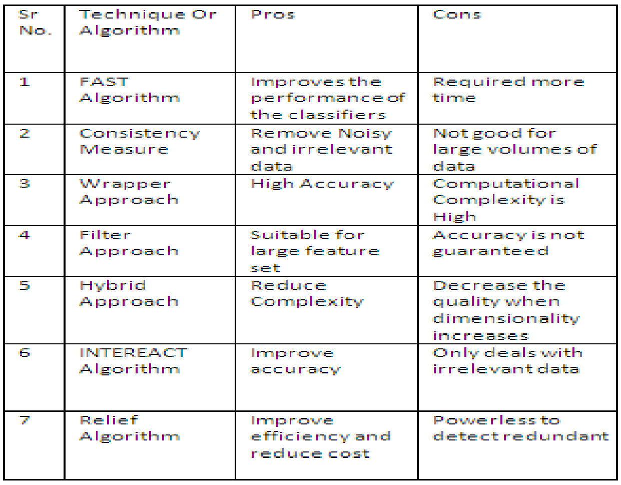
D. Subset Selection

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 selection methods must be used to obtain features that are highly correlated with the class, yet uncorrelated with each other. A novel method is proposed which can efficiently and effectively deal with both irrelevant and redundant features, and obtains a good feature subset[12].

Algorithm Analysis

The FAST algorithm contains four phases. (i) Irrelevant feature removal (ii) Distributed Clustering (iii) Redundant Feature removal (iv) Subset selection. The modified FAST algorithm is shown that represents the reduction in time to match and retrieve the data search process. This algorithm when used in various classifiers enhances their prediction capability and increases search performance[13].

Comparison of Different Algorithms/Techniques



Sr No.	Technique Or Algorithm	Pros	Cons
1	FAST Algorithm	Improves the performance of the classifiers	Required more time
2	Consistency Measure	Remove Noisy and irrelevant data	Not good for large volumes of data
3	Wrapper Approach	High Accuracy	Computational Complexity is High
4	Filter Approach	Suitable for large feature set	Accuracy is not guaranteed
5	Hybrid Approach	Reduce Complexity	Decrease the quality when dimensionality increases
6	INTEREACT Algorithm	Improve accuracy	Only deals with irrelevant data
7	Relief Algorithm	Improve efficiency and reduce cost	Powerless to detect redundant

5.CONCLUSION

Feature selection method is an efficient way to improve the accuracy of classifiers, dimensionality reduction, removing both irrelevant and redundant data. In this paper, we have made a comparative study of various feature selection methods and algorithms. In this paper we study The most important clustering algorithms are reviewed, separated in four broad categories: distance-based, density-based, model-based and grid based.

REFERENCES

[1] Qinbao Song, Jingjie Ni, and Guangtao Wang, “A Fast Clustering- Based Feature Subset Selection Algorithm for High-Dimensional Data,” IEEE Transaction on Knowledge and Data, Engineering,Vol. 25, No. 1, January 2013.

[2] C.Krier, D.Francois, F. Rossi, and M. Verleysen, “Feature Clustering and Mutual Information for the Selection of Variables in Spectral Data,” Proc. European Symp. Artificial Neural Networks Advances in Computational Intelligence and Learning, pp. 157-162, 2007.

[3] R. Butterworth, G. Piatetsky-Shapiro, and D.A. Simovici, “On Feature Selection through Clustering,” Proc. IEEE Fifth Int’l Conf. Data Mining, pp. 581-584, 2005.

[4] Butterworth R., Piatetsky-Shapiro G. and Simovici D.A., On Feature Se- lection through Clustering, In Proceedings of the Fifth IEEE international Conference on Data Mining, pp 581-584, 2005.

[5] Cardie, C., Using decision trees to improve case-based learning, In Pro- ceedings of Tenth International Conference on Machine Learning, pp 25-32, 1993.

[6] Chanda P., Cho Y., Zhang A. and Ramanathan M., Mining of Attribute Interactions Using Information Theoretic Metrics, In Proceedings of IEEE international Conference on Data Mining Workshops, pp 350-355, 2009.

[7] Yu L. and Liu H., Feature selection for high-dimensional data: aFast correlation-based filter solution, in Proceedings of 20th International Conferenceon Machine Leaning, 20(2), pp 856-863, 2003.

[8] Kohavi R. and John G.H., Wrappers for feature subset selection, Artif.Intell., 97(1-2), pp 273-324, 1997.

[9] Fleuret F., Fast binary feature selection with conditional mutual Information, Journal of Machine Learning Research, 5, pp 1531- 1555, 2004.

[10] Yu L. and Liu H., Efficient feature selection via analysis of relevance and redundancy, Journal of Machine Learning Research,10(5), pp 1205-1224,2004

[11] P. Soucy, G.W. Mineau, A simple feature selection method for text classification, in: Proceedings of IJCAI-01, Seattle, WA, 2001, pp. 897–903.

[12] 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.

[13] Zheng Zhao and Huan Liu in “Searching for Interacting Features”, ijcai07

[14] S.Swetha1, A.Harpika “A Novel Feature Subset Algorithm For High Dimensional Data” IJRRECS/October 2013/Volume-1/Issue-6/1295-1300 ISSN 2321-5461.

[15] S. Chikhi and S. Benhammada, “ReliefMSS: A Variation on a Feature Ranking Relieff Algorithm,” Int’l J. Business Intelligence and Data Mining, vol. 4, nos. 3/4, pp. 375-390, 2009.

[16] W. Cohen, “Fast Effective Rule Induction,” Proc. 12th Int’l Conf. Machine Learning (ICML ’95), pp. 115-123, 1995.

[17] M. Dash and H. Liu, “Feature Selection for Classification,” Intelligent Data Analysis, vol. 1, no. 3, pp.131-156, 1997.

[18] M. Dash, H. Liu, and H. Motoda, “Consistency Based Feature Selection,” Proc. Fourth Pacific Asia Conf.Knowledge Discovery and Data Mining, pp. 98-109, 2000.

[19] Battiti R., Using mutual information for selecting features in supervised neural net learning, IEEE Transactions on Neural Networks, 5(4), pp 537-550, 1994.

[20] Bell D.A. and Wang, H., A formalism for relevance and its application in feature subset selection, Machine Learning, 41(2), pp 175-195, 2000.

[21] Biesiada J. and Duch W., Features election for high-dimensional datała Pearson redundancy based filter, Advances in Soft Computing, 45, pp 242C249, 2008.