A FAST FUZZY MINIMUM SPANNING TREE CLUSTERING BASED FEATURE SUBSET SELECTION ALGORITHM FOR HIGH-DIMENSIONAL DATA

V.Deepa[#1], T. Deepa[*2]

Research scholar, Department of Computer Science, Sri Ramakrishna College of Arts and Science for women, Affiliated to Bharathiar University, Coimbatore, India

[1]deepu.msccs@gmail.com

Assistant Professor, Department of Computer Science, Sri Ramakrishna College of Arts and Science for women, Affiliated to Bharathiar University, Coimbatore, India

[2]DeepaRaman12@gmail.com

ABSTRACT: **Feature selection roles plays major important concern to identify features in various fields such as medical engineering, research and development .Those selected may be estimated from both the effectiveness and helpfulness point of examination. The usefulness is connected to the value of the separation of features. Several work have been concerned in earlier work to overcome the problem of best feature subset selection .In this paper proposed a novel fast clustering along with fuzzy membership function .It involves major three steps 1) the creation of the Minimum spanning tree for features; 2) divide the tree and select feature based on highest fuzzy membership results based cluster -based method for representation precise feature selection results for classification. Proposed Fuzzy clustering based feature selection calculate a novel illustration of data that optimally preserves feature selection results and form a new clusters for classification result analysis. The arithmetical evaluation confirms the classification accuracy of proposed Fuzzy FAST higher in terms of parameters classification accuracy and computational time.**

Index Terms: Fuzzy clustering, Minimum spanning tree, irrelevant and redundant features, and fast clustering-based feature selection algorithm, feature subset selection, classification, Feature selection (FS).

1. *INTRODUCTION*

Many factors concern the achievement of data mining algorithms on a specified task. The feature of the information is individual issue if information is inappropriate or the data is noisy and unpredictable, then information detection throughout training is more complex. So very efficient methods required to remove noisy data, it is known as feature selection, it removes irrelevant and unpredictable features in the dataset.

Many attribute based feature selection methods overcome the problem of best feature selection in earlier work to distinguish each features [1]. The investigate process is shared with a characteristic effectiveness estimator in order to estimate the best values of features. When the estimation of the selected features with related to learning algorithms is measured as fine it show the way to a huge.

Feature selection method is commonly second-hand as a preprocessing stage to machine learning. Consequently, FS becomes very necessary for machine learning tasks while in front of elevated dimensional information at the near time.

Though, this tendency of extent on together size and dimensionality furthermore poses severe challenges to FS methods. A number of the new investigate efforts in Fs methods have been overcome these high dimensional data of example [2-4]. In this paper major work is concerned about efficient feature selection algorithm for high dimensional data.

In this work propose a Fast Fuzzy clustering bAsed feature Selection algoriThm (FFAST). The FFAST algorithm works in three steps. In the first step, creation of MST for that the point of view the correlation and estimation of features evaluated based on fuzzy concepts and then proceed second step, based on this results the MST are created and group the highest membership features into similar group for classification purposes .Features in dissimilar clusters are comparatively self-determining; the Fuzzy clustering based strategy of FFAST has a elevated prospect of construct a subset of helpful and self-governing features.

2. *BACKGROUND STUDY*

FS methods majorly classified into two ways: filter model or the wrapper model [3]. The initial filtering methods depend on individual characteristics of training information to select most important features without learning process;

consequently it does not succeed to some preconception of a learning algorithm. The second methods wrapper algorithm necessitates individual prearranged knowledge algorithm in FS and uses its presentation to estimate classification accuracy. However, the major problem of wrapper methods is restricted and computationally complex for larger and high dimensional dataset .The major problem of the Filter methods computationally well suited, but the accurateness of the machine learning methods such as SVM, NN, Decision tree is not guaranteed [5-7] ,these methods are known as embedded methods .

Combination of filter and wrapper methods have been also anticipated in previous works [8-11] to reduce subset of features and improves accuracy of classification for high dimensional data. Although the wrapping methods is computationally expensive and be likely to overfit on little training sets [12-13].

A number of obtainable algorithms exposed effective in eliminating together inappropriate and unnecessary features incorporate the consistency measure [13] and the correlation measure [14-15]. Both correlation measure and consistency measure evaluates to find most important features to separate classes as constantly as the complete position of features preserve. An abnormality is specific as two instances comprise the identical feature values but dissimilar course group labels. The major problem of these methods is computationally not well suited in terms of dimensionality.

Of the numerous FSS algorithms, several can successfully remove inappropriate features however not succeed to handle unnecessary features [16-17], however a number of others can remove the inappropriate whereas pleasing care of the unnecessary features [18].

3. *PROPOSED METHODOLOGY*

In proposed system for gene data or the classification of data for larger data, irrelevant features and important features are removed, along with unnecessary features; strictly influence the accurateness of the knowledge technology [15]. Furthermore, novel algorithms [19] which can proficiently and successfully arrangement with together inappropriate and unnecessary features, and acquire a high-quality feature subset. The inappropriate feature elimination is easy once the accurate significance assess is distinct or elected, while the unnecessary feature removal is a small piece of complicated. In our proposed FFAST algorithm, it absorb 1) the structure of the MST based on the correlation and relevance that related to Fuzzy concept for every feature for high dimensional dataset 2) the separation of the data into dissimilar supposition of fuzzy ; and 3) the collection of illustration features beginning the clusters with maximum membership values in the fuzzy determine.

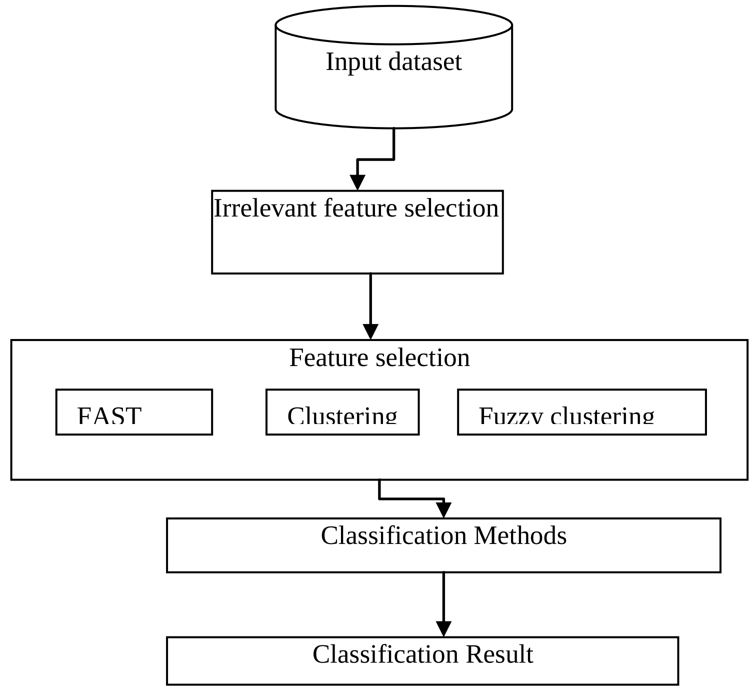

FIGURE 1: PROPOSED WORK REPRESENTATION

The FAST clustering algorithm starts with the dataset D with m features $F = \{F_1, F_2 \dots F_m\}$ and class C, to select features first need to compute $T-Relevance\ SU(F_i, C)$ that related to classes $F_i (1 \le i \le m)$ in the first step of the MST algorithm for feature selection. If established relevance value is measured based on predefined threshold value θ, it is defined $F' = \{F'_1, F'_2, \dots\dots, F'_k\}$ $(k \le m)$. In the succeeding step, principal compute the $F - Correlation\ SU(F'_i, F'_j)$ value for both pair of features F'_i and F'_j $(F'_i, F'_j \in F' \wedge i \neq j)$.Then, presentation features F'_i and F'_j as vertices and $SU(F'_i, F'_j)$ $(i \neq j)$ as the weight of the edge among vertices in the graph G F'_i and F'_j, where $V = \{F'_i | F'_i \in F' \wedge i \in [1, k]\}$
and $E = \{(F'_i, F'_j) | (F'_i, F'_j\ F' \wedge i, j \in [1, k] \wedge i \neq j)\}$.

As symmetric improbability is symmetric supplementary the $F - Correlation\ SU(F'_i, F'_j)$ is symmetric as well, in G graph . The total graph G duplicates the association amongst each and every one of the desired features in the dataset. If the graph consists of k vertices and $k(k-1)/2$ edges with dissimilar weights are robustly interwoven. Consequently for graph G, construct MST, which join every one vertices such with the intention of the addition of the weights of the edges is the smallest amount, by means of the well known Prim algorithm. After construction the MST, in the third step, principal eliminate the edges E, whose values of functions are smaller than both of the $T - Relevance\ SU(F_i^{'}, C)$ and $SU(F_j^{'}, C)$, from the MST.

Consider two different tree from MST as T_1 and T_2. The results of feature selection vertices in every features if connected to concluding trees to be $V(T)$, have the belongings that for every one pair of vertices $SU(F_i^{'}, F_j^{'}) \geq SU(F_i^{'}, C) \vee SU(F_i^{'}, F_j^{'}) \geq SU(F_j^{'}, C)$constantly hold accurate designed for each $F_i \in S(i \neq j)$, then F_i are unnecessary features with value to the known F_j. For every cluster $V(T_j)$ decide a representative feature F_R^j whose T-relevance SU(F_R^j, C)is the maximum.

For Irrelevant Feature Removal

Step 1: From the given dataset D set number of features $1\ to\ m$ and division label $C.D = (F_1, F_2, \ldots \ldots \ldots . F_M, C)$ Where $F = (F_1, F_2, \ldots \ldots \ldots . F_M)$

Step 2: Begin T-relevance $SU\ (Fi;\ C)$ value for each feature in the given dataset.

Step 3: If established relevance value is measured based on predefined threshold value θ, it is defined $F^{'} = \{F_1^{'}, F_2^{'}, \ldots \ldots, F_k^{'}\}$ $(k \leq m)$.

//Minimum Spanning Tress Construction

Step 4: MST graph G is a whole graph and establish the importance of features from step 3.

Step 5: Then evaluate F-correlation value for correlation of both whose features selected after threshold satisfies from step 3 F-Correlation $SU(F_i^{'}, F_j^{'})$ value for each pair of features $F_i^{'}$ and $F_j^{'}$

Step 6: if is better add the above features to graph and assign weight to those features and new weighted absolute graph G (V, E) is constructed.

Step 7: MST is marked by means of prism algorithm for graph G. Every one of vertices are connected accordingly that the adding together of the weights of the edges is the lowly quantity, using prism algorithm.

Tree Partition and Representative Feature Selection

Step 8: Remove the edges if the weight are smaller than threshold ideals both of the T-Relevance SU($(F_i^{'}, C)$ and SU($(F_j^{'}, C)$ from the MST.

Step 9: A afforest is achieved, for every tree in the afforest correspond to a group that is indicated as $V(T_j)$

// Redundancy are Eliminated

Step 10: For every cluster $V(T_j)$ decide a representative feature F_R^j whose T-relevance SU(F_R^j, C) is the maximum.

Step 11: All $F_R^j (j = 1, \ldots [Forest])$ consist of the concluding feature subset UF_R^j .

In this work additionally add fuzzy based Minimum spanning tree algorithm to find the T-relavance. In subtractive clustering method, features are considered as the candidates for cluster centers. Consider a collection of *features* T-relevance $SU\ (Fi;\ C)$ importance for each feature in the given dataset. The features in the high dimensional dataset are rescaled into [0, 1] in every measurement. If is better add the above features to graph is equal to threshold θ_i and assign weight to those features and new weighted absolute graph G (V, E) is constructed. The feature subset is defined as $F^{'} = \{F_1^{'}, F_2^{'} \ldots . F_k^{'}\}$ $(K \leq m)$. Define a measure of the potential of $thershold\ \theta$ as below:

$$\theta_i = \sum_{j=1}^{m} \exp\ (-\alpha ||F_i - F_j||^2), \alpha = 4/r_a^2$$

Where $||.||$represent the Euclidean distance, and r_a is a positive constant. The constant r_a is successfully a regularize radius important a region in the feature subset points this radius have small manipulate on the possible. After the possible of highest features in the dataset is selected as the first cluster center. Let F_i^* be the position of the primary cluster central point and θ_i^* be its possible importance. Revise the potential of each feature point $F_i$by the formula as follows:

$$\theta_i = \theta_i - \theta_i^* \exp\left(-\beta ||F_i - F_j^*||^2\right), \beta = \frac{4}{r_b^2}, r_b = 1.25 r_a$$

Take away a quantity of potential $\theta_i^*$from every one characteristic feature space points of its distance beginning the initial feature subset selection cluster center. The feature points near the first cluster center determination include significantly concentrated prospective, and consequently would be improbable to be preferred as the subsequently cluster center to choose features. The constant r_b is a radius important the neighborhood which determination include considerable decrease in prospective. In the succeeding step, principal compute the $F - Correlation\ SU(F_i', F_j')$ value for both pair of features F_i' and F_j' $(F_i', F_j' \in F' \wedge i \neq j)$.Then, presentation features F_i' and F_j' as vertices and $SU(F_i', F_j')$ $(i \neq j)$ as the weight of the edge among vertices in the graph G F_i' and F_j', where $V = \{F_i' | F_i' \in F' \wedge i \in [1, k]\}$ and $E = \{(F_i', F_j') | (F_i', F_j' F' \wedge i, j \in [1, k] \wedge i \neq j)\}$. As symmetric improbability is symmetric supplementary the $F - Correlation\ SU(F_i', F_j')$ is symmetric as well, in G graph .

$$\theta_{if} = \theta_{is} - \theta_{is}^{*} \exp\left(-\beta \|F_i' - F_j'\|^2\right)$$

Therefore for graph G, build an MST, which attach every one vertices such that the sum of the weights of the edges is the smallest amount, using the well known Prim algorithm . The weight of edge (F_i', F_j') is $F - Correlation\ SU(F_i', F_j')$. After building the MST, in the third step, we first take away the edges $E = \{(F_i', F_j') | (F_i', F_j' F' \wedge i, j \in [1, k] \wedge i \neq j)\}$, whose weights are smaller than both of the $T - Relevance\ SU(F_i', C)$ and $SU(F_j', C)$, from the MST. From θ_{if} & θ_i from a cluster result using T relevance and $F - Correlation\ SU(F_i', F_j')$.Each deletion results in two disconnected trees T_1 and T_2.Each cluster center may be translated into a fuzzy rule. Suppose cluster center F_{ij}^{*} was found in the group of data for class c_m , this cluster center provides the rule: *Rule i : if* F_1 *is* A_{i1} *and* F_2 *is* A_{i2} *and ... then class cm*

$$A_{ij}(F_j) = \exp\left(-\frac{1}{2}\left(F_j - \frac{F_{ij}^{*}}{\sigma_{ij}}\right)^2, \sigma_{ij}^2 = \frac{1}{2\alpha}, \alpha = \frac{4}{r_a^2}\right.$$

where F_{ij}^{*} is the feature of the j-th element of F_{ij}^{*}, and r_a is a positive constant to select features. The degree of completion of each features in the rule is followed as

$$\mu_i = \exp\left(-\alpha \|F - F_i^{*}\|^2\right)$$

Consider two different tree from MST as T_1 and T_2. The results of feature selection vertices in every features if connected to concluding trees to be $V(T)$, have the belongings that for every one pair of vertices $SU(F_i', F_j') \geq SU(F_i', C) \vee SU(F_i', F_j') \geq SU(F_j', C)$ constantly hold accurate designed for each $F_i \in S(i \neq j)$, then F_i are unnecessary features with value to the known F_j. For every cluster $V(T_j)$ decide a representative feature F_R^j whose T-relevance $SU(F_R^j, C)$ is the maximum. The stage classification, the resulting of the rule through the maximum quantity of achievement is preferred to be the yield class of the classifier $E = \frac{1}{2}(1 - \mu_{c,max} + \mu_{-c,max})^2$

For Irrelevant Feature Removal

Step 1: From the given dataset D set number of features $1\ to\ m$ and division label $C.D = (F_1, F_2, \dots\dots\dots, F_M, C)$ Where $F = (F_1, F_2, \dots\dots\dots, F_M)$

Step 2: Begin T-relevance $SU\ (Fi;\ C)$ value for each feature in the given dataset.

Step 3: If established relevance value is measured based on predefined threshold value θ, it is defined $F' = \{F_1', F_2', \dots\dots, F_k'\}$ $(k \leq m)$.

Step 4: Define a measure of the potential of $thershold\ \theta$ as below:

$$\theta_i = \sum_{j=1}^{m} \exp\left(-\alpha \|F_i - F_j\|^2\right), \alpha = 4/r_a^2$$

Where $\|.\|$ denotes the Euclidean distance to measure features , and r_a is a optimistic constant.

Step 5: Revise the potential of each feature point F_i by the formula as follows:

$$\theta_i = \theta_i - \theta_i^{*} \exp\left(-\beta \|F_i - F_j^{*}\|^2\right), \beta = \frac{4}{r_b^2}, r_b = 1.25 r_a$$

//Minimum Spanning Tress Construction

Step 6: MST graph G is a whole graph and establish the importance of features from step 3.

Step 7: Then evaluate F-correlation value for correlation of both whose features selected after threshold satisfies from step 3 F-Correlation $SU(F_i', F_j')$ value for each pair of features F_i' and F_j'

Step 8: Remove the edges if the weight are smaller than threshold ideals both of the T-Relevance SU((F_i', C) and SU((F_j', C) from the MST.

$$\theta_{if} = \theta_{is} - \theta_{is}^{*} \exp\left(-\beta \|F_i' - F_j'\|^2\right)$$

Step 9: if is better add the above features to graph and assign weight to those features and new weighted absolute graph G (V, E) is constructed.

Tree Partition and Representative Feature Selection

Step 10: Remove the edges if the weight are smaller than threshold ideals both of the T-Relevance SU((F_i', C) and SU((F_j', C) from the MST.

Remove the edges whose weights are smaller than both of the T-Relevance SU((F_i', C) and SU((F_j', C) from the MST. Suppose cluster center F_{ij}^{*} was establish in the cluster of data for class c_m , this cluster center provides the rule: *Rule i : if* F_1 *is* A_{i1} *and* F_2 *is* A_{i2} *and ... then class* c_m

$$A_{ij}(F_j) = \exp\left(-\frac{1}{2}\left(F_j - \frac{F_{ij}^{*}}{\sigma_{ij}}\right)^2\right), \sigma_{ij}^2 = \frac{1}{2\alpha}, \alpha = \frac{4}{r_a^2}$$

Step 11: A afforest is achieved, for every tree in the afforest correspond to a group that is indicated as $V(T_j)$

// Redundancy are Eliminated

Step 12: For every cluster $V(T_j)$ decide a representative feature F_R^j whose T-relevance SU(F_R^j, C) is the maximum.

Step 13: All $F_R^j\ (j = 1, \ldots [Forest])$ consist of the concluding feature subset UF_R^j .

4. *EXPERIMENTAL RESULTS*

For the purpose examination of classification result for selected best feature subset for high dimensional data, several statically analysis of the results, performed a via Demsar [20] and Garcia and Herrerato [21] to multiple data sets . In this paper proposed work has been implemented and tested on three widely presented different micro array data sets namely Lymphoma, Colon and Leukemia data sets. The descriptions of the each datasets are follows:

Lymphoma Dataset

Lymphoma dataset under the category of cancer disease dataset relies on genes type's .Diffuse large B-cell lymphomas (DLBCL) and follicular lymphomas (FL) are two major important category of gene B cell in lymphoma dataset in extremely dissimilar clinical appearance. So these two types considered as major classes for classification of selected B cell gene type features with fuzzy FAST clustering feature subset selection algorithm.

Colon Dataset

This dataset is parallel to the mushroom gene expression data set. For every example in the colon dataset, it is point toward whether it approaches starting tumor biopsy or not. It is used in numerous diverse investigate papers on gene expression data. The major diagnostic classes in colon dataset are colon normal and colon cancer.

Leukemia dataset

The leukemia datasets consists of information related to leukemia data for patient samples .It consists of measurement of two important cells lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) samples from bone substance and tangential blood [WW, 5]. These consists of two major classes namely ALL and AML.

Datasets	Samples	Number of classes
Lymphoma	56	2
Colon	61	2
Leukemia	63	2

Table 4.1. Dataset description

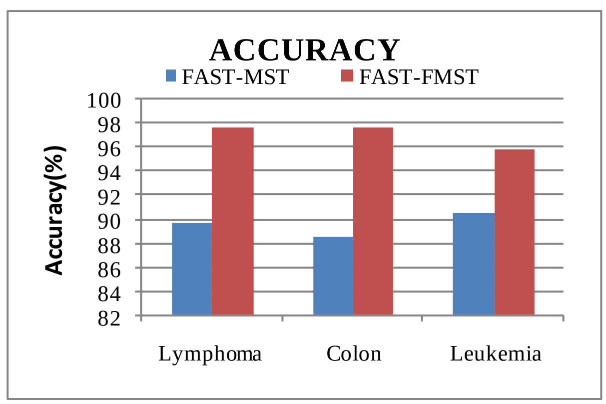

Figure 2. Accuracy comparison of dataset

DATASETS	FAST-MST	FAST-MST FUZZY CLUSTERING
LYMPHOMA	89.7	97.72

COLON	88.45	97.67
LEUKEMIA	90.56	95.81

Table 2: Accuracy results for three data sets

The above Figure 2 shows the result of classification accuracy of three dataset for FAST and FFAST feature selection results, values are tabulated in Table 2.

5. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a novel fuzzy similarity measure for FAST clustering-based feature subset selection. The algorithm removes unrelated features for clustering based on fuzzy concept and building of MST system for cluster structure based on fuzzy results. Using this new demonstration, enhanced classification rates and makes easy enlarge the rapidity of the classification procedure. In the proposed system selected features are independent to each other for separate clusters in the fuzzy membership function. Each cluster from fuzzy concept is considered as individual features characteristic features and consequently dimensionality is significantly concentrated. It shows proposed FFAST have higher classification accuracy than FAST results. For future work we plan to consider various correlation similarity measures, and study a number of prescribed belongings of feature space.

REFERENCES

1. Avrim Blum and Pat Langley, “Selection of relevant features and examples in machine learning," Artificial Intelligence, vol. 97, no. 1-2, pp. 245-271, 1997.
2. Liu, H., Hussain, F., Tan, C., & Dash, M. (2002a). Discretization: An enabling technique. Data Mining and Knowledge Discovery, 6, 393–423.
3. Das, S. (2001). Filters, wrappers and a boosting-based hybrid for feature selection. Proceedings of the Eighteenth International Conference on Machine Learning (pp. 74–81).
4. Xing, E., Jordan, M., & Karp, R. (2001). Feature selection for high-dimensional genomic microarray data. Proceedings of the Eighteenth International Conference on Machine Learning (pp. 601–608).
5. I. Guyon and A. Elisseeff, “An Introduction to Variable and Feature Selection,” J. Machine Learning Research, vol 3, pp. 1157- 1182, 2003.
6. T.M. Mitchell, “Generalization as Search,” Artificial Intelligence, vol. 18, no. 2, pp. 203-226, 1982.
7. J. Souza, “Feature Selection with a General Hybrid Algorithm,” PhD dissertation, Univ. of Ottawa, 2004.
8. P. Langley, “Selection of Relevant Features in Machine Learning,” Proc. AAAI Fall Symp. Relevance, pp. 1-5, 1994.
9. S. Das, “Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection,” Proc. 18th Int’l Conf. Machine Learning, pp. 74- 81, 2001.
10. 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.
11. J. Yu, S.S.R. Abidi, and P.H. Artes, “A Hybrid Feature Selection Strategy for Image Defining Features: Towards Interpretation of Optic Nerve Images,” Proc. Int’l Conf. Machine Learning and Cybernetics, vol. 8, pp. 5127-5132, 2005.
12. L. Yu and H. Liu, “Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution,” Proc. 20th Int’l Conf. Machine Leaning, vol. 20, no. 2, pp. 856-863, 2003.
13. Dash, M., Liu, H., & Motoda, H. (2000). Consistency based feature selection. Proceedings of the Fourth Pacific Asia Conference on Knowledge Discovery and Data Mining (pp. 98–109). Springer-Verlag.
14. Hall, M. (1999). Correlation based feature selection for machine learning. Doctoral dissertation, University of Waikato, Dept. of Computer Science.
15. Hall, M. (2000). Correlation-based feature selection for discrete and numeric class machine learning. Proceedings of the Seventeenth International Conference on Machine Learning (pp. 359–366).
16. G. Forman, “An Extensive Empirical Study of Feature Selection Metrics for Text Classification,” J. Machine Learning Research, vol. 3, pp. 1289-1305, 2003.
17. M. Scherf and W. Brauer, “Feature Selection by Means of a Feature Weighting Approach,”

Technical Report FKI-221-97, Institut fur Informatik, Technische Universitat Munchen, 1997.

18. H. Liu and R. Setiono, "A Probabilistic Approach to Feature Selection: A Filter Solution," Proc. 13th Int'l Conf. Machine Learning, pp. 319-327, 1996.

19. M.A. Hall and L.A. Smith, "Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper," Proc. 12th Int'l Florida Artificial Intelligence Research Soc. Conf., pp. 235-239, 1999.

20. J. Demsar, "Statistical Comparison of Classifiers over Multiple Data Sets," J. Machine Learning Res., vol. 7, pp. 1-30, 2006.

21. S. Garcia and F. Herrera, "An Extension on Statistical Comparisons of Classifiers over Multiple Data Sets for All Pairwise Comparisons," J. Machine Learning Res., vol. 9, pp. 2677-2694, 2008.

