IJICTR
International Journal of Information & Communication Technology Research

Volume 11- Number 4 – Autumn 2019 (40 -47)

A Text Classification Method Based on Combination of Information Gain and Graph Clustering

[1]Alireza Abdollahpouri[*], [2]Shadi Rahimi, [3]Fatemeh Zamani, [4]Parham Moradi
Department of Computer Engineering
University of Kurdistan
Sanandaj, Iran

E-mails: [1,4]{abdollahpouri, p.moradi}@uok.ac.ir, [2,3]{rahimishadi4, fateme.zamani1994}@gmail.com

Received: 20 April 2019 - Accepted: 7 August 2019

Abstract—**Text classification has a wide range of applications such as: spam filtering, automated indexing of scientific articles, identifying the genre of documents, news monitoring, and so on. Text datasets usually contain much irrelevant and noisy information which eventually reduces the efficiency and cost of their classification. Therefore, for effective text classification, feature selection methods are widely used to handle the high dimensionality of data. In this paper, a novel feature selection method based on the combination of information gain and FAST algorithm is proposed. In our proposed method, at first, the information gain is calculated for the features and those with higher information gain are selected. The FAST algorithm is then used on the selected features which uses graph-theoretic clustering methods. To evaluate the performance of the proposed method, we carry out experiments on three text datasets and compare our algorithm with several feature selection techniques. The results confirm that the proposed method produces smaller feature subset in shorter time. In addition, the evaluation of a K-nearest neighborhood classifier on validation data show that, the novel algorithm gives higher classification accuracy.**

Keywords-Feature selection, Information gain, text categorization, FAST algorithm.

I. INTRODUCTION

The goal of text classification is to categorize a document or text into predetermined classes based on the terms of the text. Text categorization is a well-studied problem. A main difficulty of text classification is that often text dataset has a lot of words which increases the computational complexity of text categorization and may results of low accuracy of classification, because of irrelevant and redundant terms in feature space. As a solution to this problem, feature selection techniques are used.

Feature selection is a process that selects a subset from basic feature set based on some feature importance measure.

Lewis and Ringutte [1] used the information gain criterion for feature selection in text dataset. Wiener et al. [2] applied mutual information and chi-square to select features. Yang [3] and Schutze et al. [4] used PCA to find orthogonal dimensions in the vector space of texts.

Hierarchical clustering has been widely used for word selection in the context of text classification (e.g. [5],

[*] Corresponding Author

[6], [7]). Distributional clustering has also been used for grouping the words. It can be performed based on the participation of the word in particular grammatical relations with other words [5], or based on the distribution of class labels associated with each word [6]. Since distributed clustering of words is agglomerative and has high computational cost, Dhillon et al. [7] proposed a new information-theoretic divisive algorithm for word clustering.

Butterworth et al. in [8], proposed a method to cluster the features using a special metric of Barthelemy-Montjardet distance. They used dendrogram of the resulting cluster hierarchy to choose the most relevant attributes. Unfortunately, the feature subset identified by the cluster evaluation measure based on Barthelemy-Montjardet distance, does not allow the classifier to improve the original performance accuracy. Furthermore, even compared with other feature selection methods, the obtained accuracy is lower. The FAST algorithm introduced by Song et al. [9], uses graph clustering for feature selection. This 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 classes 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. FAST algorithm use a method based on minimum spanning tree (MST) to cluster the features. But it does not assume that data points are grouped around centres or separated by a regular geometric curve [9].

Sabbah et al. [10] presented the Support Vector Machine based Feature Ranking Method (SVM-FRM) in which the weighting and ranking of features are based on the SVM learning algorithm. After that, they applied hybridization techniques to enhance the efficiency of SVM-FRM method in some experimental situations.

Rehamn et al. in [11] introduced a new feature ranking metric, namely normalized difference measure (NDM) which considers the relative document frequencies of a term in both positive and negative classes while determining the rank of a term.

In [12], the authors provide an in-depth analysis on feature selection step for text classification, and propose a novel strategy for selecting the features automatically. They formulated the feature selection process as a multiple objectives optimization problem, and identified the best number of selected features for each document automatically, rather than determining a fixed threshold to optimize the overall classification accuracy for different categories.

In this paper, information gain measure and FAST algorithm are combined to produce an appropriate feature subset for text datasets.

The purpose of this combination is two-fold. Firstly, FAST has a high probability of producing a subset of useful and independent features because of its clustering-based strategy. Secondly, an advantage of information gain is that due to the factor –p.log(p) in the entropy definition, leafs with a small number of instances are assigned less weight. Therefore, this type of combination leads to producing smaller feature subset in a very lower time in comparison with the original FAST algorithm.

In the first step, the proposed method calculates information gain of features, and then removes lower values features. Thereafter, the FAST algorithm is applied on these features to select the final feature subset.

The rest of this paper is organized as follows: in Section II, short background information about information gain (IG) is given. The proposed method is then discussed in Section III. Section IV is devoted to experimental results. Finally, in Section V, we outline the main conclusions.

II. BACKGROUND AND RELATED WORKS

A. Feature selection

Feature selection methods can be classified into the filter, wrapper, and hybrid approaches (see Fig. 1). Filter methods use an information theoretic criterion to evaluate the goodness of a feature or a set of features. In the wrapper approach, a classifier is used and trained to evaluate a set of prominent features [13]. However, due to a learning model being involved in the searching process of the wrapper approach, these methods often suffer from high computational cost and loss of generality. The hybrid approach takes the advantages of both filter and wrapper approaches. Filter methods are fast enough and their results do not rely on a specific classifier and thus are appropriate for real-world applications.

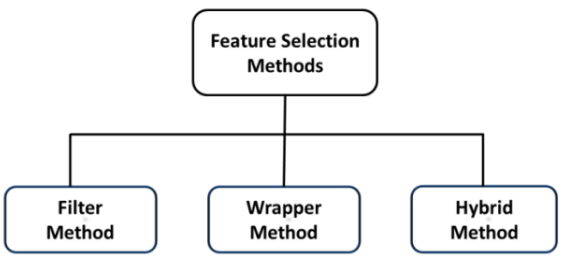

Figure 1. Feature selection methods.

B. Information Gain

Information gain (IG) is a feature evaluation method which is used in the field of machine learning. In feature selection, IG measures the amount of information provided by the features for the target feature. This measure is defined as (1):

$$IG(t) = -\sum_{i=1}^{m} P_r(C_i)\log_2 P_r(C_i) + P_r(t)\sum_{i=1}^{m} P_r(c_i|t)\log_2 P_r(c_i|t_i) + P_r(\bar{t})\sum_{i=1}^{m} P_r(c_i|\bar{t})\log_2 P_r(c_i|\bar{t}) \quad (1)$$

where, *C* is a set of categories. For each unique term, we calculate the *IG* measure, and remove the terms from the feature space whose *IG* was less than a predefined threshold.

FAST algorithm composed of two components: *irrelevant feature removal* and *redundant feature elimination*. This algorithm involves 1) removing irrelevant features, 2) constructing a minimum spanning tree from relative ones, and 3) partitioning the MST and selecting representative features.

C. Related works

Previous studies have shown that filter-based methods are much successful than others. From one point of view, the filter-based methods are categorized into univariate and multivariate methods. Univariate methods used an information theoretic criterion to evaluate the relevancy of features to the target class. Up to this time, several univariate criteria have been proposed in the literature such as Information Gain (IG) [14], Mutual Information (MI) [15, 16], Document Frequency (DF) [17], Term strength (TS), Bi-normal-Separation (BNS) [18], Odds Ratio (OR) [13], Relative Discrimination Criterion (RDC) [19], Fisher Score (FS) [20], and Laplacian Score (LS) [21]. Univariate methods are effective to evaluate features, considering their relevance to the target class. However, they ignore the correlation between features, and these methods cannot identify the redundant features. Multivariate methods consider both the relevancy of features with the target class and the correlation between selected features in their ranking processes. There are some multivariate methods, including minimal redundancy maximal relevance (mRMR) [22], Relevance redundancy feature selection (RRFS) [23], MIFS [24], Normalized mutual information feature selection (NMIFS) [25], MIFS-U [26], Unsupervised feature selection based on Ant Colony Optimization (UFSACO) [27], and Multivariate RDC (MRDC) [28]. All these methods identify prominent features by optimizing a single objective function. From the other point of view, filter-based feature selection methods can be categorized into ranking-based and subset selection-based methods [29]. Ranking-based methods first assign a relevance value to each feature using a univariate or a multivariate criterion, and then sort the features and select those of the top high scores. For example, in [19] an efficient univariate criterion, called RDC, is proposed for assigning a rank value for each term in the text classification task. RDC assigns high scores to those terms that appear frequently in a specific class. In [11], a text specific criterion, called Normalized Difference Measure (NDM), is proposed which takes into account the relative document frequencies. Some univariate methods such as IG [14], MI [15, 16], DF [17], TS BNS [18], OR [13], RDC [19], FS [20], SU [30], and LS [21] are also categorized as ranking-based methods.

TABLE I. META -HEURISTIC BASED FEATURE SELECTION METHODS.

Methods	Type	Search Method	Application
MRDC[28]	Filter/SSB	Multivariate Greedy	Textual
RRFS[23]	Filter	Multivariate Greedy	Numeric
mRMR[22]	Filter	Multivariate Greedy	Textual/Numeric
RRFSACO[35]	Filter/SSB	Multivariate ACO	Numeric
GCACO[33]	Filter/SSB	Multivariate ACO	Numeric
RDC[19]	Filter-RB	Univariate	Textual
DFS[36]	Filter-RB	Univariate	Textual
NDM[11]	Filter-RB	Univariate	Textual
F-Score[20]	Filter-RB	Univariate	Textual/Numeric
Gini-Index[37]	Filter-RB	Univariate	Textual/Numeric
MI[15]	Filter-RB	Univariate	Textual/Numeric
LS[21]	Filter-RB	Univariate	Numeric
DF[17]	Filter-RB	Univariate	Textual
IG[14]	Filter-RB	Univariate	Textual/Numeric
BNS[18]	Filter-RB	Univariate	Textual
CHI[38]	Filter-RB	Univariate	Textual/Numeric
GR[39]	Filter-RB	Univariate	Numeric

Although ranking-based methods require low computational resources, all these methods consider only the relevancy of the features and neglect the redundancy with others. Identifying a set of optimal feature subset that results in building a learning model with maximum accuracy is an NP-hard problem [29]. To overcome this issue, the subset selection-based methods seek to find a near optimal feature set by applying some heuristic or meta-heuristic methods. For example, Relevance redundancy feature selection (RRFS) [23], Mutual information feature selector (MIFS) [24], Normalized mutual information feature selection (NMIFS) [25], MIFS-U [26], MIFS-ND[31], JMIM [32], Online Streaming Feature selection based on Mutual Information (OSFMI) and MRDC [28] use sequential forward or backward selection as type of greedy search strategy, and thus they easily trap into a local optima. To solve this issue, some researchers

have focused on applying nature-inspired methods such as Genetic Algorithm (GA), Particle Swarm Optimization (PSO), and Ant Colony Optimization (ACO) to find a near optimal subset.

Many of the existing methods consider the feature selection task as a single-objective optimization problem. For example, the author of [27], proposed an unsupervised filter method for feature selection. Their method called UFSACO employed ACO to search through the feature space and proposed a feature counting metric to evaluate a subset of features. The same authors extended this work and proposed RRFSACO [35] which considers both relevancy and redundancy of features in the search process of ants in ACO. In GCACO [33] and MGCACO [34] the graph clustering with ACO was used for feature selection. All these methods use some specific information theoretic criterion to evaluate a set of features. The difference between these methods is based on different evaluation functions and different search strategies. Most of these methods use various types of relevancy metrics and ignore the redundancy between features. Although these methods are successful in finding valuable feature sets, they often have some major issues.

Table 1 summarizes the main properties of meta-heuristic based feature selection methods. This table reports three main properties including, feature selection type, search method, and application domain.

III. PROPOSED METHOD

We present a novel feature selection technique based on Information gain criteria and FAST algorithm. Before presenting our method, we describe concepts of relevant feature and redundant feature. Irrelevant features, along with redundant features, severely affect the accuracy of the learning machines [40]. Thus, feature subset selection method must be able to identify and remove irrelevant and redundant information as much as possible. Meanwhile, a good feature subset contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other. Traditional definitions of relevant and redundant features are defined as follows [41]. Suppose F to be the full set of features, $F_i \in F$ be a feature, $S_i = F - \{F_i\}$ and $S'_i \subseteq S_i$. Let s'_i be a value-assignment of all features in S'_i, f_i a value-assignment of feature F_i, and c a value-assignment of target concept C. relevant and redundant feature definition can be formalized as follows:

Definition 1 (Relevant Feature). F_i is relevant to the target concept C if and only if there exists some s'_i, f_i, and c, such that, for probability $p(S'_i = s'_i, F_i = f_i) > 0$

$$p(C = c \mid S'_i = s'_i, F_i = f_i) \neq p(C = c \mid S'_i = s'_i)$$

Otherwise, feature F_i is an *irrelevant feature*.

Definition 2 (Markov Blanket). Given a feature $F_i \in F$, let $M_i \subset F$ ($F_i \notin M_i$), M_i is said to be a Markov blanket for F_i if and only if:

$$p(F - M_i - \{F_i\}, C \mid F_i, M_i) = p(F - M_i - \{F_i\}, C \mid M_i).$$

Definition 3 (Redundant Feature). Let S be a set of features, a feature in S is redundant if and only if it has a Markov Blanket within S.

Theory of relevant feature and redundant feature is in terms of feature correlation and feature-target correlation.
The symmetric uncertainty (SU) [42] 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 for classification by a number of researchers (e.g. Yu and Liu [31], [43], Zhao and Liu[44], [45]). In FAST algorithm, authors choose symmetric uncertainty as the measure of correlation between either two features or a feature and the target feature. The symmetric uncertainty is defined as follows [42]:

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

where, H(X) is the entropy of feature X and is defined as follows:

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \quad (3)$$

Gain (X|Y) is the additional information about random variable Y that provided by X. It is defined as (4):

$$Gain(X|Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \quad (4)$$

where, H (X|Y) is the entropy of X after observing variable Y. This measure is defined by (5):

$$H(X|Y) = -\sum_j P(y_j) \sum_i P(x_i|y_j) \log_2 P(x_i|y_j) \quad (5)$$

SU values normalized to the range [0, 1], the value 1 of SU(X, Y) shows that X and Y are completely dependent. The value of 0 indicates that X and Y are independent.

Definition 4 (F_redundancy). Let $S = \{F_1, F_2, \ldots, F_i, \ldots, F_k<|F|\}$ be a cluster of features. If $\exists F_j \in S$, $SU(F_j, C) \geq SU(F_i, C) \wedge SU(F_i, F_j) > SU(F_i, C)$ is always corrected for each $F_i \in S$ ($i \neq j$), then F_i is redundant feature with respect to the given F_j (i.e., each F_i is an F-redundancy).

The proposed method consists of two main steps as follows (see Figure 2):

1. In the first step, information gain is calculated for each term using (1). Then, the p percent of features with higher information gain is selected.
2. The second step of the proposed technique, related to FAST algorithm [9], consists of three main steps, as follows:
   i) In this phase, for each selected term in step 1, correlation to target feature using (2), is calculated. Then features whose

this measure is less than a predefined threshold, are removed from the feature space.

ii) In this phase, correlation between all terms is calculated (using (2)), the weighted complete graph using this terms and their correlation is built. Then, minimum spanning tree of this graph is constructed.

iii) In last part of our method (which is same as FAST algorithm), the MST is partitioned into sub-trees by eliminating edges that their weights are lower than correlation of both nodes with target feature. This work redounds to rising clusters or trees that their features are redundant according to definition (4). Hence selecting of only one feature from each cluster for forming feature subset, is efficient. Thus, we select the features by highest correlation with target from each cluster as representative feature of clusters.

Pseudo-code of our proposed method (IG + FAST) can be formalized as follows:

Algorithm 1 IG + FAST

// Part 1

1) Calculate information gain of all terms using (1)
2) Select p% of best features in terms of IG value

// Part 2

3) For each selected term T in previous phase, calculate SU(T, C) using (2)
4) Remove features with SU lower than predefined threshold

// Part 3

5) Calculate correlation between features that are selected in step 6, using (2)
6) Create a complete graph that weight of edge of between F_i' and F_j' equals to $SU(F_i', F_j')$
7) Create MST of this graph by Prime algorithm

// Part 4

8) Delete those E_{ij} edges that $SU(F_i', F_j') < SU(F_i', C) \wedge SU(F_i', F_j') < SU(F_j', C)$
9) For each tree or cluster select the feature with maximum SU

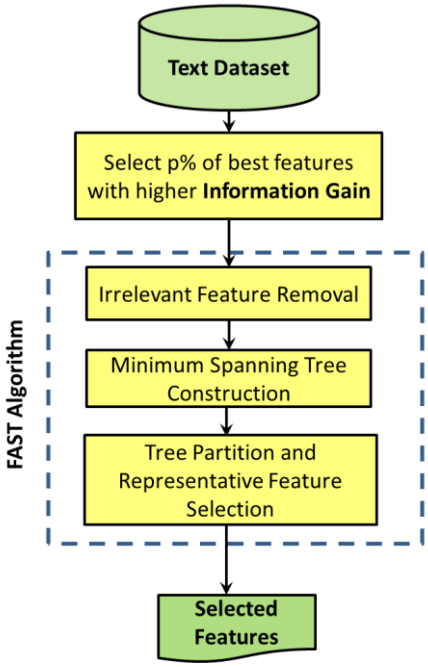

Figure 2. Framework of the proposed method.

IV. PEFORMANCE EVALUATION

In this section, we study the effectiveness of our approach and compare the results obtained by IG+FAST w.r.t. similar algorithms namely, FAST, DF + FAST, MI + FAST and CHI + FAST.

A. *Description of Data Sets*

We applied our proposed method on the following datasets to evaluate and compare its performance.

1) tr23.wc: multi- class (1-of-n) text dataset denoted by George Forman. This data set contains 204 instances and its feature space has 5832 dimension, the number of classes is 6.

2) fbis.wc: multi- class text which contains 2463 instances and its feature space has 2000 dimension, and the number of classes is 17.

3) tr21.wc: multi- class (1-of-n) text dataset denoted by George Forma. This data set containing 336 instances and its feature space has 7902 dimension, the number of classes is 6.

B. *Evaluation environment and conditions*

The proposed method was implemented in Matlab, on a computer with Intel® Core™ i7 CPU 2.67 GHz and 8 GB of memory.

To have a more precise evaluation of the performance of our proposed algorithm, all approaches have been run ten times and the average values are reported. In DF+FAST method, we combined

document frequency with FAST instead of information gain, and in MI+FAST we combined mutual information measure with FAST and in CHI + FAST, we used chi-square measure for combining with FAST algorithm. We used KNN classifier to classify datasets before and after feature selection for all different types of feature selection algorithms. The relevant threshold for all datasets was set to 0.04, and *p* parameter was set as 20 percent.

We evaluate the performance of the feature subset selection algorithms, by means of the following three metrics, 1) the proportion of selected features 2) the time to obtain the feature subset, 3) the classification accuracy. The proportion of selected features is the ratio of the number of features selected by a feature selection algorithm to the original number of features of a data set.

C. Results and Analysis

1) Proportion of selected features: Table 2 records the proportion of the five feature selection algorithms for each data set. In general, all of them achieve significant reduction of feature space.

TABLE II. PROPORTION OF SELECTED FEATURES OF THE FIVE ALGORITHMS.

Data set	IG + FAST	FAST	DF + FAST	MI + FAST	CHI + FAST
tr23.wc	0.07	0.15	0.09	0.05	0.1
fbis.wc	0.19	0.8	0.13	0.2	0.2
tr21.wc	0.07	0.1	0.05	0.05	0.15

From this, we observe that, generally all five algorithms achieve significant reduction of dimensionality by selecting only a small portion of the basic features. Our proposed algorithm produces smaller feature subset compared to FAST algorithm. MI + FAST for tr23.wc and tr21.wc has best performance in reducing feature space.

2) Run-time of algorithm: Table 3 shows the run time of the five feature selection algorithms for each dataset.

TABLE III. RUN TIME (IN SECONDS) OF THE FIVE ALGORITHMS.

Data set	IG + FAST	FAST	DF + FAST	MI + FAST	CHI + FAST
tr23.wc	827	1670	675	857	801
fbis.wc	170	514	85	185	168
tr21.wc	375	1254	336	984	979

It can be observed that our proposed algorithm compared to FAST algorithm, produces feature subset in a shorter time. DF + FAST for all dataset have shortest run-time, because time complexity of document frequency measure is low.

3) Classification accuracy of KNN classifier:

Figure 3 shows accuarcy of KNN classifier on the three data sets before and after each feature selection technique.

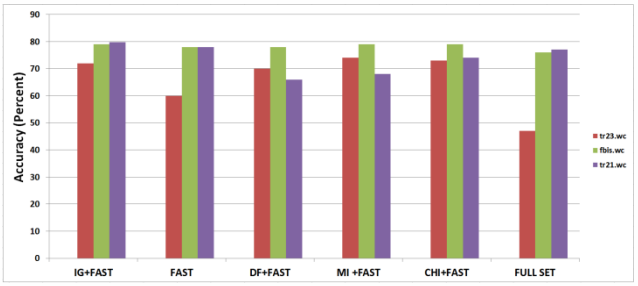

Figure 3. Classification accuracy of KNN classifier.

It is obviously concluded that, for tr23.wc data set, the proposed algorithm after MI + FAST has best accuracy. For fbis.wc data set, IG + FAST, MI + FAST and CHI + FAST have best accuracy. For tr21.wc data set after FAST, IG + FAST have best classification accuracy. For tr21 data set compared with original data, IG + FAST algorithm increases the classification accuracy by 1.8 percent.

Finally, in terms of classification accuracy, from Figure 3 we observe that in general, our proposed method obtains the rank of 1, and CHI+FAST ranks 2; Although MI+FAST provides the best accuracy for tr23.wc data set, it stands in the 3rd place when considering overall accuracy in all datasets.

V. CONCLUSION

In this paper, a new feature selection technique, which combines information gain measure and FAST algorithm, has been introduced. We have compared the performance of the proposed algorithm with four feature selection methods, namely, FAST, DF + FAST, MI+FAST and CHI+FAST on three text data sets from the three aspects of the proportion of selected features, runtime and classification accuracy of KNN classifier. Results show that IG + FAST algorithm improves classification accuracy, and in a shorter time, produces smaller feature subset, as well. As the future work, one could explore different types of correlation measures, and study some formal properties of feature space. Meanwhile, some more Meta-heuristic based feature selection methods can be investigated for text classification.

REFERENCES

[1] D. D. Lewis, and M. Ringuette, "A comparison of two learning algorithms for text categorization," In Third annual symposium on document analysis and information retrieval, pp. 81-93, 1994.
[2] E. Wiener, J. O. Pedersen, and A. S. Weigend, "A neural network approach to topic spotting," Document Analysis and Information Retrieval, 1995.
[3] Y. Yang, "Noise reduction in a statistical approach to text categorization," presented at the Research and Development in Information Retrieval, 1995.

[4] H. Schutze, D. A. Hull, and J. O. Pedersen, "A Comparison of classifiers and document representation for the routing problem," presented at the Research and Development in Information Retrieval, 1995.
[5] F. Pereira, N. Tishby, and L. Lee, "Distributional Clustering of English Words," presented at the Meeting on Assoc. for Computational Linguistics, 1993.
[6] L. D. Baker and A. K. McCallum, "Distributional Clustering of Words for Text Classification," presented at the Research and Development in information Retrieval, 1998.
[7] I. S. Dhillon, S. Mallela, and R. Kumar, "A Divisive Information Theoretic Feature Clustering Algorithm for Text Classification," Machine Learning Research, vol. 3, pp. 1265-1287, 2003.
[8] R. Butterworth, G. Piatetsky-Shapiro, and D. A. Simovici, "On Feature Selection through Clustering," presented at the Data Mining, 2005.
[9] Q. Song, J. Ni, and G. Wang, "A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data," IEEE Transactions on Knowledge and Data Engineering, vol. 25, 2013.
[10] T. Sabbah, M. Ayyash, and M. Ashraf, "Hybrid support vector machine based feature selection method for text classification," Int. Arab J. Inf. Technol., 15(3A), pp. 599-609, 2018.
[11] A.Rehman, K.Javed, and H. A.Babri, "Feature selection based on a normalized difference measure for text classification," Information Processing and Management, 53(2), pp. 473-489, 2017.
[12] C. Huang, J. Zhu, Y. Liang, M.Yang, G. P. C. Fung, and J. Luo, "An efficient automatic multiple objectives optimization feature selection strategy for internet text classification, " International Journal of Machine Learning and Cybernetics, pp.1-13, 2018.
[13] B. Agarwal and N. Mittal, "Text classification using machine learning methods-a survey," in Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012, 2014, pp. 701-709.
[14] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in ICML, 1997, pp. 412-420.
[15] Y. Xu, G. J. Jones, J. Li, B. Wang, and C. Sun, "A study on mutual information-based feature selection for text categorization," Journal of Computational Information Systems, vol. 3, pp. 1007-1012, 2007.
[16] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," bioinformatics, vol. 23, pp. 2507-2517, 2007.
[17] L. Liu, J. Kang, J. Yu, and Z. Wang, "A comparative study on unsupervised feature selection methods for text clustering," in Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference on, 2005, pp. 597-601.
[18] G. Forman, "An extensive empirical study of feature selection metrics for text classification," The Journal of machine learning research, vol. 3, pp. 1289-1305, 2003.
[19] A. Rehman, K. Javed, H. A. Babri, and M. Saeed, " Relative discrimination criterion–A novel feature ranking method for text data," Expert Systems with Applications, vol. 42, pp. 3670-3681, 2015.
[20] Q. Gu, Z. Li, and J. Han, "Correlated multi-label feature selection," in Proceedin gs of the 20th ACM international conference on Information and knowledge management, 2011, pp. 1087-1096.
[21] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in Advances in neural information processing systems, 2005, pp. 507-514.
[22] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 27, pp. 1226-1238, 2005.
[23] A. J. Ferreira and M. A. Figueiredo, "An unsupervised approach to feature discretization and selection," Pattern Recognition, vol. 45, pp. 3048-3060, 2012.
[24] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," Neural Networks, IEEE Transactions on, vol. 5, pp. 537-550, 1994.
[25] P. A. Estévez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," Neural Networks, IEEE Transactions on, vol. 20, pp. 189-201, 2009.
[26] N. Kwak and C.-H. Choi, "Input feature selection for classification problems," Neural Networks, IEEE Transactions on, vol. 13, pp. 143-159, 2002.
[27] S. Tabakhi, P. Moradi, and F. Akhlaghian, "An unsupervised feature selection algorithm based on ant colony optimization," Engineering Applications of Artificial Intelligence, vol. 32, pp. 112-123, 2014.
[28] M. Labani, P. Moradi, F. Ahmadizar, and M. Jalili, "A novel multivariate filter based feature selection method for text classification problems," Engineering Applications of Artificial Intelligence, vol. 70, pp. 25-37, 2018.
[29] F. Zarisfi Kermani, E. Eslami, and F. Sadeghi, "Global Filter–Wrapper method based on class-dependent correlation for text classification," Engineering Applications of Artificial Intelligence, vol. 85, pp. 619-633, 2019/10/01/ 2019.
[30] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in ICML, 2003, pp. 856-863.
[31] N. Hoque, D. K. Bhattacharyya, and J. K. Kalita, "MIFS-ND: A mutual information-based feature selection method," Expert Systems with Applications, vol. 41, pp. 6371-6385, 2014/10/15/ 2014.
[32] M. Bennasar, Y. Hicks, and R. Setchi, "Feature selection using Joint Mutual Information Maximisation," Expert Systems with Applications, vol. 42, pp. 8520-8532, 2015/12/01/ 2015.
[33] P. Moradi and M. Rostami, "Integration of graph clustering with ant colony optimization for feature selection," Knowledge-Based Systems, vol. 84, pp. 144-161, 2015/08/01/ 2015.
[34] H. Ghimatgar, K. Kazemi, M. S. Helfroush, and A. Aarabi, "An improved feature selection algorithm based on graph clustering and ant colony optimization," Knowledge-Based Systems, vol. 159, pp. 270-285, 2018/11/01/ 2018.
[35] S. Tabakhi and P. Moradi, "Relevance–redundancy feature selection based on ant colony optimization," Pattern Recognition, vol. 48, pp. 2798-2811, 2015/09/01/ 2015.
[36] A. K. Uysal and S. Gunal, "A novel probabilistic feature selection method for text classification," Knowledge-Based Systems, vol. 36, pp. 226-235, 2012/12/01/ 2012.
[37] W. Shang, H. Huang, H. Zhu, Y. Lin, Y. Qu, and Z. Wang, "A novel feature selection algorithm for text categorization," Expert Systems with Applications, vol. 33, pp. 1-5, 2007.
[38] F. Sebastiani, "Machine learning in automated text categorization," ACM computing surveys (CSUR), vol. 34, pp. 1-47, 2002.
[39] A. G. Karegowda, A. Manjunath, and M. Jayaram, "Comparative study of attribute selection using gain ratio and correlation based feature selection," International Journal of Information Technology and Knowledge Management, vol. 2, pp. 271-277, 2010.
[40] A.GowriDurga, A. Priya, "Feature Subset Selection Algorithm for High Dimensional Data using Fast Clustering Method, " IJCAT International Journal of Computing and Technology, 1(2), 2014.
[41] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant Features and the Subset Selection Problem," presented at the Machine Learning, 1994.
[42] S. I. Ali, W. Shahzad, "A feature subset selection method based on symmetric uncertainty and ant colony optimization, " In Emerging Technologies (ICET), International Conference on, pp. 1-6, 2012.
[43] L. Yu and H. Liu, "Efficient Feature Selection via Analysis of Relevance and Redundancy," Machine Learning Research, vol. 10, pp. 1205-1224, 2004
[44] Z. Zhao and H. Liu, "Searching for Interacting Features," presented at the Artificial Intelligence, 2007.
[45] Z. Zhao and H. Liu, "Searching for Interacting Features in Subset Selection," Intelligent Data Analysis, vol. 13, pp. 207-228, 2009.

Alireza Abdollahpouri is an Associate professor of Computer Networks at the Department of Computer Engineering, University, Kurdistan, Sanandaj, Iran. He has obtained Ph.D. (Computer Networks) in 2012 from University of Hamburg, Germany. He received the B.Sc. and M. Sc. degrees both in Computer Engineering from Isfahan University of Technology and Amirkabir University of Technology, respectively. He was working at Telecommunication Company of Tehran from 2001 until 2005 as a system programmer. He joined the University of Kurdistan as a faculty member in 2005. His main research interests are in the field of IPTV over Wireless Networks, Performance Evaluation and Modeling the Network Systems.

Shadi Rahimi has M.Sc. in Artificial Intelligence from University of Kurdistan, Sanandaj, Iran. She received the B. Sc. degree in Computer Engineering from university of Kurdistan, Sanandaj, Iran. Her research interests include Community Detection in Social Networks, Text Mining, Data Mining, and Machine Learning

Fatemeh Zamani has M.Sc. in Artificial Intelligence from University of Kurdistan, Sanandaj, Iran. She received the B.Sc. degree in Computer Networks Engineering from university of Kurdistan. Her research interests include Fog Computing, Text Mining, and Machine Learning

Parham Moradi received Ph.D. degree in Computer Science from Amir kabir University of Technology in March 2011. Moreover, He received M.Sc. and B.Sc. degree in Software Engineering and Computer Science from Amirkabir University of Technology, Tehran, Iran, in 1998 and 2005, respectively. He conducted a part of his Ph.D. research work in the Laboratory of Nonlinear Systems, EPFL (Ecole Polytechnique Federal de Lausanne), Lausanne, Switzerland, from September 2009 to March 2010. Currently he works as an Associate Professor in the Department of Computer Engineering and Information Technology, University of Kurdistan, Sanandaj, Iran. His current research areas include Machine Learning, Feature Selection, Social Network Analysis, Data Mining and Recommender Systems.