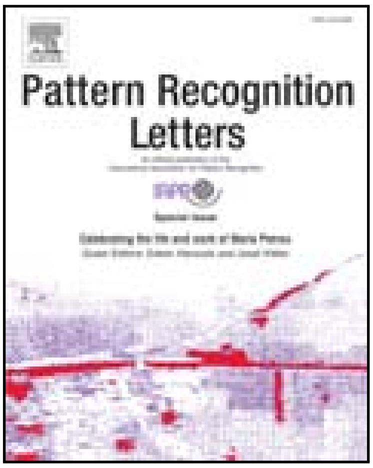

Feature selection considering the composition of feature relevancy

Wanfu Gao, Liang Hu, Ping Zhang, Jialong He

PII: S0167-8655(18)30228-9
DOI: 10.1016/j.patrec.2018.06.005
Reference: PATREC 7199

To appear in: *Pattern Recognition Letters*

Received date: 30 July 2017
Revised date: 12 February 2018
Accepted date: 4 June 2018

Please cite this article as: Wanfu Gao, Liang Hu, Ping Zhang, Jialong He, Feature selection considering the composition of feature relevancy, *Pattern Recognition Letters* (2018), doi: 10.1016/j.patrec.2018.06.005

Highlights

- A novel feature selection method based on information theory is proposed.
- The composition of feature relevancy is taken into account.
- Our method maximizes new information while minimizing redundancy information.
- Our method outperforms five competing methods in terms of average accuracy.
- The proposed method achieves the highest accuracy.

Pattern Recognition Letters
journal homepage: www.elsevier.com

Feature selection considering the composition of feature relevancy

Wanfu Gao[a], Liang Hu[a,**], Ping Zhang[b], Jialong He[a,c]

[a]College of Computer Science and Technology, Jilin University, Changchun 130012, China
[b]College of Software, Jilin University, Changchun 130012, China
[c]School of Mechanical Science and Engineering, Jilin University, Changchun 130012, China

ABSTRACT

Feature selection plays a critical role in classification problems. Feature selection methods intend to retain relevant features and eliminate redundant features. This work focuses on feature selection methods based on information theory. By analyzing the composition of feature relevancy, we believe that a good feature selection method should maximize new classification information while minimizing feature redundancy. Therefore, a novel feature selection method named Composition of Feature Relevancy (CFR) is proposed. To evaluate CFR, we conduct experiments on eight real-world data sets and two different classifiers (Naïve-Bayes and Support Vector Machine). Our method outperforms five other competing methods in terms of average classification accuracy and highest classification accuracy.

Keywords: Feature selection; Information theory; Classification; Composition of feature relevancy

1. Introduction

Feature selection plays a critical role in classification problems, especially for data sets that have many features. These features can be roughly divided into two groups: relevant features and redundant features. To select relevant features and eliminate redundant features, many feature selection methods have been proposed (Bolón-Canedo et al., 2014; Li et al., 2016).

With respect to different selection strategies, feature selection methods can be broadly categorized into three models: filter models, wrapper models and embedded models (Wang et al., 2017b; Senawi et al., 2017; Hu et al., 2018). Wrappers select a feature subset based on a specified classifier such as Naïve-Bayes (NB) and Support Vector Machine (SVM) classifiers. Wrappers perform well in classification, however, they suffer from a risk of over-fitting. Similarly, embedded models select the best feature subset during the learning process of a specified learning algorithm. Embedded and wrapper models are both dependent on a specified classifier; thus, they are computationally expensive. Filters are independent of a specified classifier, and have low computational cost. In this work, we focus on filters.

Information theory is an important metric that is extensively used in filter models (Das and Das, 2017; Bennasar et al., 2013). Many feature selection methods based on information theory have been proposed. These methods can be generally categorized into two groups based on different evaluation criteria (Wang et al., 2017a): minimizing feature redundancy or maximizing new classification information. Nevertheless, these two groups of methods can be transformed into each other under certain conditions (Brown et al., 2012).

To select relevant features and eliminate redundant features, we begin by making an analysis of new classification information. We discover that new classification information consists of feature relevancy and feature redundancy. Further, through equation transformation, we can obtain the composition of feature relevancy. By analyzing the composition of feature relevancy, we propose a novel feature selection method based on information theory named Composition of Feature Relevancy (CFR).

To evaluate our method, CFR is compared to five competing methods on eight real-world data sets. We obtain average classification accuracies on two different classifiers, Naïve-Bayes (NB) and Support Vector Machine with RBF kernel (SVM). CFR achieves good classification performance. Furthermore,

[**]Corresponding author: Liang Hu

our method obtains the highest accuracies on eight data sets using each feature selection method.

The rest of this paper is organized as follows. In Section 2, we provide some preliminaries for this work. In Section 3, we review related work. In Section 4, we propose a novel feature selection method. In Section 5, we show experimental results and discussion. We conclude this work in Section 6 and give a plan for future action.

2. Preliminaries

In this section, we briefly introduce information theory (Cover and Thomas, 2012) that includes mutual information, conditional mutual information, joint mutual information and interaction information. Consider three discrete random variables X, Y and Z. The mutual information between X and Y is defined as:

$$I(X;Y) = H(Y) - H(Y|X) \tag{1}$$

where $H(Y)$ and $H(Y|X)$ are entropy and conditional entropy. Entropy describes the uncertainty of a random variable, and conditional entropy describes the amount of uncertainty left when another variable is introduced. Therefore, the mutual information represents the uncertainty reduced when another variable is introduced; it also indicates the amount of information that both variables share.

Another important concept of information theory is conditional mutual information, which is expressed as follows:

$$I(X;Y|Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z) \tag{2}$$

Conditional mutual information represents the amount of information between X and Y when Z is introduced. Similarly, joint mutual information between (X,Y) and Z is defined as:

$$I(X,Y;Z) = I(X;Z) + I(Y;Z|X) \tag{3}$$

Joint mutual information indicates the mutual information between (X,Y) and Z.

Interaction information is used to measure feature redundancy in the process of feature selection. It is defined as:

$$I(X;Y;Z) = I(X;Z) + I(X;Y) - I(X;Y,Z) \tag{4}$$

Interaction information can be positive, zero or negative. It is positive when the variables Y and Z can provide the same information; it is negative when Y and Z provide more information than they provide individually; it is zero when Y and Z are independent in the context of X. Eq. (4) can be rewritten as follows (Wang et al., 2017a):

$$I(X;Y;Z) = I(X;Y) - I(X;Y|Z) \tag{5}$$

In other sources, interaction information is defined using the opposite sign in Eq. (5); however, this is for mathematical convenience and does not change the final conclusion.

3. Related work

Many feature selection methods based on information theory have been proposed. The most direct method for evaluating the significance of features is to measure the mutual information between features and classes, that is, Mutual Information Maximization(MIM) (Lewis, 1992). Although MIM has low time complexity, it ignores the feature redundancy between features.

Battiti (Battiti, 1994) proposes Mutual Information Feature Selection (MIFS) to address this issue. The criterion of MIFS is as follows:

$$J(X_k) = I(X_k;Y) - \beta \sum_{X_j \in S} I(X_j;X_k) \tag{6}$$

where X_k represents a candidate feature, X_j is a selected feature, Y is the class and S is the subset of selected features. $J(.)$ is a feature selection criterion such that, the higher the value of $J(X_k)$, the more important the feature X_k is. The feature redundancy term $\beta \sum_{X_j \in S} I(X_j;X_k)$ is introduced by calculating the mutual information between a candidate feature and each selected feature. The parameter β is set to 1 according to the suggestion of the author in our experiments.

Unlike MIFS, Minimum-Redundancy Maximum-Relevance (mRMR) (Peng et al., 2005) sets the parameter β be the reverse of the number of selected features:

$$J(X_k) = I(X_k;Y) - \frac{1}{|S|} \sum_{X_j \in S} I(X_j;X_k) \tag{7}$$

MIFS and mRMR provide different parameters β to balance feature relevancy and feature redundancy. They both achieve good classification performance.

In the literature (Brown et al., 2012), a general framework for feature selection method based on information theory is formulated as:

$$J(X_k) = I(X_k;Y) - \beta \sum_{X_j \in S} I(X_j;X_k) + \lambda \sum_{X_j \in S} I(X_j;X_k|Y) \tag{8}$$

It indicates that MIFS and mRMR are both a variation of Formula (8) when β and λ take different values.

Joint Mutual Information (JMI) (Yang and Moody, 2000) selects the feature that maximizes the joint mutual information between candidate features and classes in the context of selected features. The criterion of JMI is as follows:

$$\begin{aligned} J(X_k) &= \sum_{X_j \in S} I(X_k, X_j; Y) && (9) \\ &= \sum_{X_j \in S} \{I(X_k;Y|X_j) + I(X_j;Y)\} && (10) \end{aligned}$$

We can discover that $\sum_{X_j \in S} I(X_j;Y)$ is regarded as a constant in the feature selection process. Therefore, Eq. (10) can be rewritten:

$$J(X_k) \simeq \sum_{X_j \in S} I(X_k;Y|X_j) \tag{11}$$

It indicates that JMI selects the feature that obtains the maximal value of new classification information. Alternatively, JMI is also a variation of Formula (8). It can be reformulated as follows:

$$J(X_k) = I(X_k;Y) - \frac{1}{|S|}\sum_{X_j \in S} I(X_j;X_k) + \frac{1}{|S|}\sum_{X_j \in S} I(X_j;X_k|Y) \tag{12}$$

β and λ are both equal to $\frac{1}{|S|}$. Like JMI, Conditional Mutual Information Maximization (CMIM) combines new classification information with the 'maximum of the minimum' criterion. It is presented as follows:

$$J(X_k) = argmax_{X_k \in F-S}(min_{X_j \in S}(I(X_k;Y|X_j))) \tag{13}$$

where F is the entire set of features.

In 2015, Zeng et al. (Zeng et al., 2015) propose a novel feature selection method named Interaction Weight based Feature Selection (IWFS) that achieves good classification performance. IWFS takes feature interdependence into account. IWFS proposes an interaction weight factor, defined as:

$$IW(X_k, X_j) = 1 + \frac{I(X_j;X_k;Y)}{H(X_k) + H(X_j)} \tag{14}$$

IW represents interaction weight. The criterion of IWFS is as follows:

$$J(X_k) = IW(X_k, X_j) + IW(X_k, X_j) * SU(X_k;Y) \tag{15}$$

$SU(X_k;Y)$ represents the normalized feature relevancy.

4. The proposed Composition of Feature Relevancy method

As mentioned in Section 1, one group of feature selection methods aims to maximize new classification information, i.e., $I(X_k;Y|X_j)$. According to Eq. (5),

$$I(X_k;Y|X_j) = I(X_k;Y) - I(X_k;Y;X_j) \tag{16}$$

The first item on the right side of Eq. (16) is feature relevancy: mutual information between a candidate feature and the class. Similarly, the second item represents feature redundancy. Eq. (16) can be rewritten as follows:

$$I(X_k;Y) = I(X_k;Y|X_j) + I(X_k;Y;X_j) \tag{17}$$

Therefore, we can conclude that feature relevancy consists of two parts: new classification information and redundant information. Feature selection methods intend to maximize the new classification information while minimizing redundant information. Intuitively, we propose a novel feature selection method based on information theory named Composition of Feature Relevancy (CFR). Its criterion is as follows:

$$J(X_k) = \sum_{X_j \in S}\{I(X_k;Y|X_j) - I(X_k;Y;X_j)\} \tag{18}$$

According to the concept of information theory in Section 2, the criterion of CFR is equivalent to the following form; the derivation process is as follows:

$$\begin{aligned} J(X_k) &= \sum_{X_j \in S}\{I(X_k;Y) - 2I(X_k;Y;X_j)\} \\ &= \sum_{X_j \in S}\{I(X_k;Y) - 2\{I(X_k;X_j) - I(X_k;X_j|Y)\}\} \\ &= |S|I(X_k;Y) - 2\sum_{X_j \in S} I(X_k;X_j) + 2\sum_{X_j \in S} I(X_k;X_j|Y) \\ &\simeq I(X_k;Y) - \frac{2}{|S|}\sum_{X_j \in S} I(X_k;X_j) + \frac{2}{|S|}\sum_{X_j \in S} I(X_k;X_j|Y) \end{aligned} \tag{19}$$

CFR can also be regarded as a specific form of Formula (8). Similar to mRMR and JMI, in CFR, the weight of the redundancy term is negatively correlated with the cardinality of the feature subset. It is helpful to reduce the effect of the redundancy term when increasing the number of selected features (Li et al., 2016). For the sake of fairness, our method and the comparison methods both employ the same greedy searching strategy, namely, Sequential Forward Search (SFS). The feature subset S starts as an empty set, then, selects one informative feature based on the feature selection method each time until the number of seleted features is larger than the user-specified threshold k.

The steps of our method are described as follows:

1. (Initialization) Set $F \leftarrow$ "Original feature set of n features", $S \leftarrow$ "empty set".
2. (Calculate mutual information between the class with each candidate feature) For each feature $X_k \in F$, calculate $I(X_k;Y)$.
3. (Select the first feature) Find the feature that maximizes $I(X_k;Y)$, $F \leftarrow F \setminus \{X_k\}$; $S \leftarrow \{X_k\}$.
4. (Greedy selection) Repeat until $|S| = k$
    (a) (Calculate conditional mutual information and interaction information) For all pairs of variables X_k and X_j such that $X_k \in F$, $X_j \in S$, calculate $I(X_k;Y|X_j)$ and $I(X_k;Y;X_j)$.
    (b) (Select the next feature) Choose the feature X_k that maximizes Formula (18).
5. The output is the set S that includes the selected features.

The method chooses the most relevant feature as the first feature. Then, interaction information and conditional mutual information are calculated to select the feature that maximizes the criterion in Formula (18). The procedure is ended when $|S| = k$.

5. Experimental results and discussion

5.1. Experimental results and analysis

In this section, we show the main experimental results involving average classification accuracies and the highest accuracies. Our method is compared with the five baselines JMI, MIFS, MIM, mRMR and IWFS on eight benchmark data sets. The characteristics of these data sets are depicted in Table 1. All the experiments were executed on an Intel Core i7 with a 3.40 GHZ processing speed and 8 GB main memory.

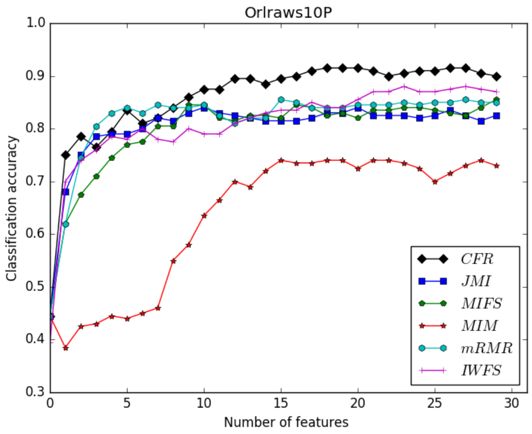

Fig. 1. Average classification accuracy achieved with Orlraws10P

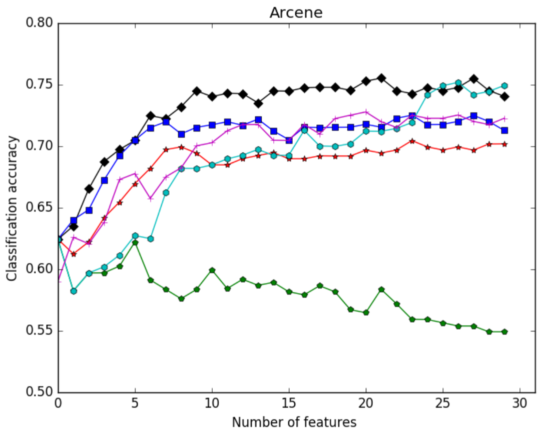

Fig. 2. Average classification accuracy achieved with Arcene

Table 1. Description of data sets

Data sets	#Instances	#Features	#Classes	Types
Arrhythmia	430	257	2	continuous
Orlraws10P	100	10304	10	continuous
TOX_171	171	5748	4	continuous
SMK_CAN_187	187	19993	2	continuous
Arcene	200	10000	2	continuous
COIL20	1440	1024	20	continuous
Isolet	1560	617	26	continuous
Landsat	6435	36	7	continuous

In Table 1, we observe that these benchmark data sets cover different number of instances, classes and features. Additionally, the features of these eight benchmark data sets are all continuous; we discretize continuous features into five bins using equal-width discretization. 10-fold cross-validation is used in this experiment. These data sets come from UCI (Lichman, 2013) and (Li et al., 2016); these data sets are also used in the literature (Aksakalli and Malekipirbazari, 2016; Zhang et al., 2017).

First, we employ NB and SVM classifiers to test the average classification accuracy. The average classification accuracies of two classifiers across 30 groups of feature subsets are recorded in Tables 2 and 3. Furthermore, a paired two-tailed t-test is conducted between CFR and other methods. The notation '+'/'-'/'=' indicates the statistically significant (at 5%) that our method wins/losses/equals over other methods. The last row (W/T/L) indicates that the number of the data set has higher (or equal, lower) accuracy than comparison methods. The bold font indicates the maximal value of the row.

From Tables 2 and 3, we find that CFR achieves the highest average accuracies with respect to two different classifiers.

Furthermore, four representative results of average classification accuracy by two classifiers are shown in Figs. 1-4. In Figs. 1-4, the x-axis indicates the first number of selected features and the y-axis is the average classification accuracy. As shown in Fig. 1, different colors and shapes indicate different feature selection methods.

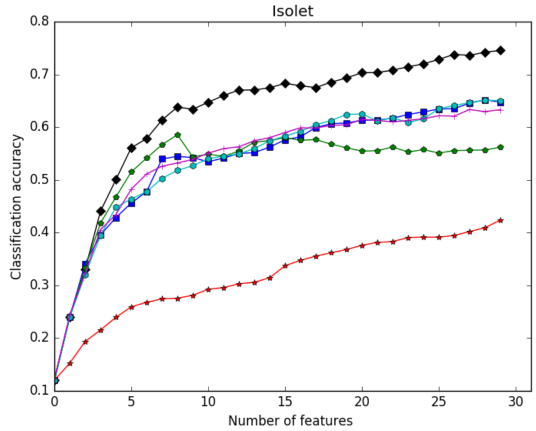

Fig. 3. Average classification accuracy achieved with Isolet

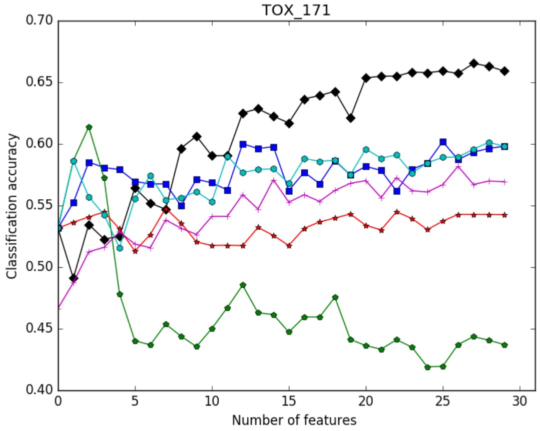

Fig. 4. Average classification accuracy achieved with TOX_171

We discover that CFR outperforms other baselines in terms of average classification accuracy of two classifiers in Figs. 1-4.

Moreover, we show the highest value of average classification accuracy of two classifiers (NB and SVM) on eight data

Table 2. Average classification accuracies (mean ± std.) of NB with p-value on eight data sets with each feature selection method.

Data Sets	JMI	MIFS	MIM	mRMR	IWFS	CFR
Arrhythmia	75.94±5.81(+)	61.97±3.53(+)	73.96±5.69(+)	69.61±3.12(+)	74.03±3.48(+)	**76.79±6.2**
Orlraws10P	71.7±5.53(+)	75.97±7.71(+)	57.73±9.06(+)	74.1±6.53(+)	82.17±9.08(=)	**82.37±8.02**
TOX_171	58.14±1.74(+)	38.59±7.54(+)	50.17±1.68(+)	54.97±1.75(+)	50.99±1.92(+)	**60.12±4.75**
SMK_CAN_187	62.44±1.03(+)	53.62±3.78(+)	61.87±1.76(+)	62±1.43(+)	63.37±3.74(+)	**66.05±2.08**
Arcene	65.93±1.3(+)	55.76±2.53(+)	65.21±0.85(+)	64.98±3.38(+)	63.92±1.78(+)	**67.26±2.22**
COIL20	68.97±12.28(+)	59.49±9.6(+)	48.58±7.56(+)	69.17±12.89(+)	67.62±12.17(+)	**71.23±14.12**
Isolet	48.34±11.08(+)	39.97±7.6(+)	28.08±6.41(+)	46.87±10.42(+)	49.37±10.37(+)	**57.9±13.48**
Landat	74.97±3.58(+)	74.78±3.58(+)	72.59±3.44(+)	74.22±3.46(+)	**76.35±3.84(-)**	75.94±3.82
Average	65.80	57.52	57.27	64.49	65.98	**69.71**
W/T/L	8/0/0	8/0/0	8/0/0	8/0/0	6/1/1	

Table 3. Average classification accuracies (mean ± std.) of SVM with p-value on eight data sets with each feature selection method.

Data Sets	JMI	MIFS	MIM	mRMR	IWFS	CFR
Arrhythmia	76.5±5.61(+)	67.96±2.1(+)	75.82±5.6(+)	76.56±4.37(=)	75.27±3.78(+)	**77.19±5.74**
Orlraws10P	88.33±9.89(=)	82.97±9.28(+)	68.87±16.78(+)	89.03±10.59(=)	79.63±9.09(+)	**89.3±10.29**
TOX_171	57.26±2.56(+)	54.38±2.79(+)	56.45±1.65(+)	59.87±3.02(+)	58.16±3.8(+)	**61.67±5.79**
SMK_CAN_187	61.13±2.25(+)	63.56±1.79(=)	62.46±1.74(+)	62.69±1.58(+)	63.56±2.82(=)	**63.59±1.3**
Arcene	75.36±4.04(+)	60.31±1.98(+)	71.39±4.24(+)	72.34±6.45(+)	75.39±5.56(+)	**78.45±4.76**
COIL20	79.89±15.67(+)	75.84±13.62(+)	64.53±13.64(+)	79.57±15.56(+)	**82.26±16.34(-)**	81.35±16.39
Isolet	59.04±13.59(+)	63.69±15.13(+)	35.28±9.08(+)	60.45±14.53(+)	58.28±13.58(+)	**66.32±16.48**
Landat	82.56±5.52(+)	82.8±5.93(+)	81.33±5.45(+)	82.29±5.71(+)	83.14±5.63(=)	**83.28±5.85**
Average	72.51	68.94	64.52	72.85	71.96	**75.14**
W/T/L	7/1/0	7/1/0	8/0/0	6/2/0	5/2/1	

Table 4. Highest value of average classification accuracy (%) of two classifiers (NB and SVM) on eight data sets by each feature selection method

Data Sets	JMI	MIFS	MIM	mRMR	IWFS	CFR
Arrhythmia	80.16(27)	70.42(3)	79.26(20)	75.27(14)	76.63(29)	**81.09**(29)
Orlraws10P	84.00(11)	85.50(30)	74.00(16)	85.50(16)	88.00(24)	**91.50**(19)
TOX_171	60.19(26)	61.37(3)	54.78(8)	60.13(29)	58.19(27)	**66.53**(28)
SMK_CAN_187	65.35(2)	64.88(2)	65.38(11)	65.09(3)	67.05(24)	**67.85**(9)
Arcene	72.52(24)	62.44(1)	70.46(24)	75.19(27)	72.79(21)	**75.54**(22)
COIL20	83.17(28)	75.12(9)	66.58(29)	84.00(29)	84.66(29)	**87.11**(29)
Isolet	65.19(29)	58.56(9)	42.37(30)	65.13(29)	63.33(28)	**74.55**(30)
Landat	81.05(30)	81.18(30)	80.81(30)	80.70(30)	81.54(28)	**81.71**(25)
Average	73.95	69.93	66.71	73.88	74.02	**78.24**

sets by each feature selection method. Experimental results are depicted in Table 4. The bold value indicates the highest accuracies of six methods. The last row (Average) shows that the average value of the highest accuracy and the numbers in the parentheses are the number of features for which the associated highest value is reached.

5.2. Complexity analysis

Suppose the number of features to be selected is k, M represents the number of instances in the data set, and the total number of features is N. The time complexity of mutual information, conditional mutual information and joint mutual information is $O(M)$ because all instances need to be examined for probability estimation. As a result, the time complexity of MIM is $O(MN)$, whereas the time complexity of CFR, JMI, MIFS, mRMR, and IWFS is $O(kMN)$.

Although MIM attains the best time complexity, the classification performance of MIM is not as good as its time complexity, which has been verified in Section 5.1. Thus, the time complexity of our method is acceptable.

6. Conclusion and future action

This paper proposes a novel feature selection method based on information theory that considers the composition of feature relevancy. The new criterion maximizes the new classification information while minimizing feature redundancy. Our method is compared to five competing feature selection methods on eight real-world data sets. Our method achieves the best classification performance in terms of average accuracy and highest accuracy. In light of the above experimental results, we conclude that our method outperforms other compared feature selection methods.

In the future, we plan to introduce a new metric into feature selection. We intend to compare a new method based on this metric with the method based on information theory.

Acknowledgments

This work is funded by: National Key R&D Plan of China under Grant No. 2017YFA0604500, National Sci-Tech Support Plan of China under Grant No. 2014BAH02F00, and by National Natural Science Foundation of China under Grant No. 61701190, and by Youth Science Foundation of Jilin Province of China under Grant No. 20160520011JH and 20180520021JH , and by Youth Sci-Tech Innovation Leader and Team Project of Jilin Province of China under Grant No. 20170519017JH, and by Key Technology Innovation Cooperation Project of Government and University for the whole Industry Demonstration under Grant No. SXGJSF2017-4.

References

Aksakalli, V., Malekipirbazari, M., 2016. Feature selection via binary simultaneous perturbation stochastic approximation. Pattern Recognition Letters 75, 41–47.

Battiti, R., 1994. Using mutual information for selecting features in supervised neural net learning. IEEE Transactions on neural networks 5, 537–550.

Bennasar, M., Setchi, R., Hicks, Y., 2013. Feature interaction maximisation. Pattern Recognition Letters 34, 1630–1635.

Bolón-Canedo, V., Sánchez-Marono, N., Alonso-Betanzos, A., Benítez, J.M., Herrera, F., 2014. A review of microarray datasets and applied feature selection methods. Information Sciences 282, 111–135.

Brown, G., Pocock, A., Zhao, M.J., Luján, M., 2012. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. Journal of machine learning research 13, 27–66.

Cover, T.M., Thomas, J.A., 2012. Elements of information theory. John Wiley & Sons.

Das, A., Das, S., 2017. Feature weighting and selection with a pareto-optimal trade-off between relevancy and redundancy. Pattern Recognition Letters 88, 12–19.

Hu, L., Gao, W., Zhao, K., Zhang, P., Wang, F., 2018. Feature selection considering two types of feature relevancy and feature interdependency. Expert Systems with Applications 93, 423–434.

Lewis, D.D., 1992. Feature selection and feature extraction for text categorization, in: Proceedings of the workshop on Speech and Natural Language, Association for Computational Linguistics. pp. 212–217.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J., Liu, H., 2016. Feature selection: A data perspective. arXiv preprint arXiv:1601.07996 .

Lichman, M., 2013. UCI machine learning repository. URL: `http://archive.ics.uci.edu/ml`.

Peng, H., Long, F., Ding, C., 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on pattern analysis and machine intelligence 27, 1226–1238.

Senawi, A., Wei, H.L., Billings, S.A., 2017. A new maximum relevance-minimum multicollinearity (mrmmc) method for feature selection and ranking. Pattern Recognition 67, 47–61.

Wang, J., Wei, J.M., Yang, Z., Wang, S.Q., 2017a. Feature selection by maximizing independent classification information. IEEE Transactions on Knowledge and Data Engineering 29, 828–841.

Wang, Y., Wang, J., Liao, H., Chen, H., 2017b. An efficient semi-supervised representatives feature selection algorithm based on information theory. Pattern Recognition 61, 511–523.

Yang, H.H., Moody, J., 2000. Data visualization and feature selection: New algorithms for nongaussian data, in: Advances in Neural Information Processing Systems, pp. 687–693.

Zeng, Z., Zhang, H., Zhang, R., Yin, C., 2015. A novel feature selection method considering feature interaction. Pattern Recognition 48, 2656–2666.

Zhang, Z., Bai, L., Liang, Y., Hancock, E., 2017. Joint hypergraph learning and sparse regression for feature selection. Pattern Recognition 63, 291–309.