2016 9th International Conference on Developments in eSystems Engineering

A Robust Feature Selection Algorithm

B. Chandra
Sprinklr and
Adjunct IIT Delhi
bchandra104@yahoo.co.in

Abstract— The paper proposes a novel feature selection algorithm based on the density function of features. Density difference between features for paired classes is used for assigning weights to the features. Existing methods select only those features that can distinguish between all classes at the same time. A new filter based feature selection method termed as Paired Class Density Difference with Inconsistency (FSDD) is proposed in this paper. It can be used for a multi-class problem. This approach selects those features that distinguish between few classes but still play an important role in classification. Inconsistency measure is used in order to remove redundancy in the selected feature set. Classification accuracy of the proposed method is compared with that obtained using existing filter based feature selection methods on UCI machine learning repository datasets and on manual segmentation data provided by NIST. Increased classification accuracy for FSDD shows that the concept of using density difference to assign feature weight is a significant contribution.

Keywords—Feature Selection, Filter approach, density function, Inconsistency measure.

I. Introduction

Feature selection algorithms play a significant role in the area of pattern classification. Some of the important applications include gene selection from microarray data [17], text categorization [16], face recognition [4], spam filter [12] and remote sensing [14]. In a pattern classification problem, all features may not be important for classification. It has been observed that pattern classification with important features gives better classification accuracy [7]. Feature selection algorithms can be categorized into two types, Filter and Wrapper approach [5]. Filter approach tries to remove the irrelevant features before they are used by the learning algorithm. The selected features are independent of the learning methods; and hence it is widely used. ReliefF [13], chi2-test [9], t-test and mRMR [3] are some of the well-known filter approach algorithms. Wrapper approach utilizes the learning algorithm as a fitness function and searches for the best features in the space of all feature subsets. This allows the use of standard optimization techniques like sequential forward selection [18] and backward feature selection [10], greedy search and randomized search like genetic algorithm [8].

Existing filter approach methods take into account all the classes for deciding the relevance of a feature. It is possible that some of the features can distinguish between a class pair and not all the classes; however they may still play a significant role in classification. A novel feature selection termed FSDD approach has been proposed in this paper that finds the relevant features that can distinguish between the class pairs. The existing feature selection methods use some statistic derived from the feature to decide the relevance of that feature but in the proposed approach the density function of the feature itself is used. The entire statistic about a feature can be derived from its density function, which shows the superiority of the proposed method over the existing methods. Inconsistency measure is used to remove redundancy in the selected features set. It has been illustrated that the features which are regarded as important by other feature selection methods are considered by FSDD In addition it is also illustrated that the features which play important role in classification but are neglected by other feature selection methods are also considered by FSDD Classification accuracy of the proposed method FSDD has been compared with existing feature selection methods on UCI machine learning datasets and on manual segmentation data from NIST. Classification accuracy of the proposed method is better cPvtompared to the existing feature selection methods.

II. Overview of Existing Methods

Various filter based feature selection algorithms have been developed in the past. Some of the existing algorithms include ReliefF [13], chi2-test [9], t-test, mRMR [3], correlation based approach to determine relevance and redundancy of a feature [6], feature selection based on consistency [2], reducing redundancy based on similarity of features [11] and Backward Elimination using Hilbert-Schmidt Independence Criterion (BAHSIC) [15]. An overview of some of the widely known feature selection methods is as follows:

A. Relief

The weight of each feature is computed as difference of weighted average of miss distances and the hit distance, with the weights being the corresponding class probabilities [13].
In this algorithm, a pattern X is selected at random. K nearest neighbors from the same class called nearest hits H and K nearest neighbors from each of the different classes, called nearest misses M are selected. The weight of each feature is updated based on H and M. If the values of a feature f in pattern X and in the patterns in H are different, it indicates that feature value within the same class is different and hence the weight of feature is decreased according to the average difference of feature value between pattern X and patterns in H.

978-1-5090-5487-9/17 $31.00 © 2017 IEEE
DOI 10.1109/DeSE.2016.70

On the other hand, if the values of a feature f in pattern X and in the patterns in M are different, it indicates that the feature has different values for different classes, hence the weight of the feature f is increased based on weighted average distance between feature value for pattern X and for patterns in M. The process of randomly selecting a pattern is repeated N times which is set by the user. Weight of each feature f is updated as follows:

Initialize weight of feature *f* to be zero.

$$W_f = 0$$

X: randomly selected instance.

H_i : Set of K nearest patterns from the same class as the class of X.

$M_{i,c}$: Set of K nearest patterns from each class c such that c is different from the class of X.

P(c): Probability of class c.

Weight of feature *f* is updated as follows:

$$W_f = W_f - \frac{1}{K}\frac{\sum_{i=1}^{K}\left|H_i^{(f)} - X^{(f)}\right|}{\max(f)-\min(f)} + \sum_{c \neq class(X)} \left\{ \frac{P(c)}{1-p(class(X))} \frac{1}{K} \frac{\sum_{i=1}^{K}\left|M_{i,c}^{(f)} - X^{(f)}\right|}{\max(f)-\min(f)} \right\} \quad (1).$$

B. Maximum Relevance Minimum Redundancy (mRMR-FCD)

Relevant features are selected based on Maximum relevance with respect to class feature and minimum redundancy with respect to other selected features. The feature having highest difference of relevance and redundancy is included in the relevant features set and the process continues till the desired number of features is selected. In mRMR-FCD, [3] F-test is used for computing relevance of a feature and for estimating redundancy, correlation is used. Initial feature is selected based on F-value and the difference of F-test value and correlation determines which feature to select next.

III. PROPOSED FEATURE SELECTION METHOD

The algorithm is composed of two parts. In the first part, initial weight of each feature is computed by Paired class density difference. The feature having greatest weight is selected. In the second part, the computed weight along with the inconsistency measure is used to include the relevant features in selected features set one by one. The feature having greatest weight- inconsistency value is selected and included in the selected feature set. The process repeats till the desired numbers of features are selected.

Motivation for using paired class information and density difference is given below.

A. Motivation for using paired class Information

For a classification problem involving more than two classes, different features are relevant for distinguishing between different pairs of classes. If we look for good features which can distinguish between all classes, then we might neglect those features which are good for distinguishing one class from other classes. In ReliefF we consider those features which can distinguish between all classes and in this process the features which distinguish only between few classes are ignored.

If a set of features is good for distinguishing between different class pairs, then increased classification accuracy can be achieved if these set of features are used for classification. Relevant information for classification can be derived even if a feature can distinguish between some pairs of classes.

B. Motivation for using Density Difference

If a feature can distinguish between two classes, then the feature values for one class will be markedly different from the values for another class. This intuition is used in many feature selection algorithms like hit and miss values in ReliefF, F-test in mRMR, t-test and Chi2 test etc.

In the proposed method, entire data is utilized directly in the form of probability density function. However in methods like relief-F, t test and Chi square, the summary statistics like Mean, Variance, and Correlation etc. are used.

C. Proposed Algorithm

Density difference between paired classes is used in the proposed algorithm for finding weights of the features. All class pairs are considered and for each pair, the weight is computed for each feature according to the overlap of the density function. Final weight of feature is the sum of weights corresponding to every pair of classes. Hence it is more inclusive as compared to Relief-F or other feature selection methods which prefer the features that can distinguish between all the classes. The feature having highest weight is included in the selected features set which is denoted by S.

For illustration, two features are considered and their density functions are shown in Figures 1 and 2 respectively. The area of shaded region shows the overlap of two density functions. It denotes the relevance of the feature for distinguishing between classes 1 and 2. For feature 1, the two density functions are highly separated, this implies that the feature values for one class are markedly different from the values of the feature for another class. So, feature 1 is highly relevant for classification. On the contrary, for feature 2, the two density functions are mostly overlapping, hence the area of shaded region is very small and the feature is not relevant for classification. If a feature can distinguish between two classes, then the corresponding area between density functions for the two classes will be around 2 and if the feature is irrelevant, then the area will be near zero.

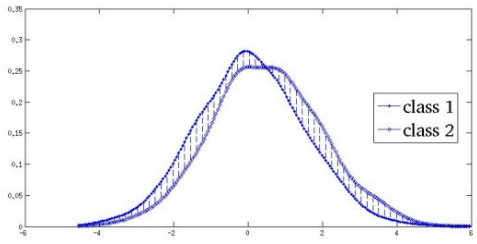

Fig.1. Probability density functions of feature 1

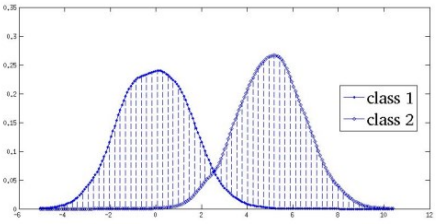

Fig. 2. Probability density functions of feature 2

For computing the density function, either parametric or non-parametric estimation can be used. Since the actual density is unknown, assuming it as normal or some other density might introduce bias in the estimation, therefore non-parametric estimation is preferred. Kernel density estimation is used to find the density of the features for each class.

Then, for every feature $f \in S$, the inconsistency of the set $\{S,f\}$ is computed. The algorithm for computing inconsistency of a set, given the class feature C, is given below.

1) Each feature in the set is discretized.

2) Inconsistency of a pattern is computed as follows,

T : count of the pattern in the dataset.

c (i): count of the pattern in the dataset, given that the corresponding class is i

m = max(c)

Inconsistency of the pattern = 1 – m/T

3) Inconsistency of Set = sum of inconsistency of all the patterns / total number of patterns

Illustration: Consider two discretized features A1, A2 and the class feature C as given in Table 1. Inconsistency of the set {A1, A2} is computed as given in Table 2.

Sum of inconsistency = 1. Inconsistency of the set = sum of inconsistency / total number of patterns = 0. 2

The feature having highest difference between weight and inconsistency is included in the selected features set. The process repeats till the desired number of features are selected.

Table 1. Features and Class Label

A1	A2	C
A	A	1
A	C	2
A	C	1
C	B	2
C	B	3

Table 2. Inconsistency calculation

Patterns	c(1)	c(2)	c(3)	m	T	Inconsistency
A A	1	0	0	1	1	0
A C	1	1	0	1	2	0.5
C B	0	1	1	1	2	0.5

IV ILLUSTRATION AND RESULTS

Detailed working of the proposed method FSDD on Seeds dataset, taken from UCI machine learning repository, having 7 features is presented in this section. This is to demonstrate that if a feature is good for distinguishing between all the classes, then the same feature can distinguish between every pair of classes and will be selected by FSDD. Hence, FSDD does not neglect those features that maybe selected by other feature selection methods. The density functions of seven features of Seeds dataset for three classes is shown in Figure 3.

Feature 1 can clearly distinguish between classes 2 and 3. For class pair (1, 3) as well as for class pair (1, 2), there is little overlap of density function but while taking the sum of density difference for all the pairs i.e., class pairs (1, 2), (1, 3) and (2, 3), feature 1 has highest value. Hence, feature 1 is selected as the topmost ranking feature by FSDD. Feature 1 is also selected by mRMR and ReliefF. This is the case where a feature that is good for distinguishing between all the classes is also good for distinguishing between all the class pairs.

FSDD selects feature 7 as the second feature. Feature 7 can clearly separate class 1 from class 2, so it complements the already selected feature 1. Selected feature set {1, 7} can clearly distinguish between class pair (2, 3) as well as between class pair (1, 2).The feature set {1, 7} provides increased classification accuracy compared to the top two features selected by mRMR, chi2 and ReliefF (see Table 4).

The rank of selected features by each method remains same for majority of iterations during 10 fold cross-validation, which is given in Table 3. The average classification accuracy obtained by Naïve Bayes classifier for 10 fold cross validation (using same feature set in every iteration of 10 fold CV) is mentioned in Table 4.

Table.3 Features selected from Seeds Dataset

rank	mRMR	chi2	Relief-F	FSDD
1	1	2	1	1
2	2	5	2	7
3	5	1	5	6
4	7	4	7	2
5	4	7	4	5
6	3	3	3	4
7	6	6	6	3

Table.4. 10 fold CV accuracy of various features set

Feature Set	mRMR	Chi2	ReliefF	FSDD
Top 1	86.67	87.62	86.67	86.67
Top 2	86.19	84.29	86.19	91.9
Top 3	85.71	85.71	85.71	94.29
Top 4	89.05	86.67	89.05	92.38
Top 5	88.09	88.09	88.09	90.47

From Table 4, it is evident that when the features selected by FSDD are used for classification, they provide better classification accuracy as compared to the features selected by other feature selection methods. It can also be observed from Table 4 that feature 7 is an important feature which can be seen in the columns corresponding to mRMR and ReliefF(in Table 4) that when feature 7 is included in selected features set, the classification accuracy drastically increases from 85.71 to 89.05. Likewise, in chi2, when feature 7 is included, again the classification accuracy in increased from 86.67% to 88.09. However FSDD selects feature 7 as the second best feature and the classification accuracy increases from 86.67% to 91.9% when feature 7 is included. Choice of feature 7 as the second best feature is ignored by other feature selection methods.

Comparative performance evaluation was also done on other benchmark datasets taken from UCI Machine Learning Repository and on manual segmentation dataset provided by NIST, Gaithersburg, Maryland. Description of datasets is given in Table 5. For the data sets in Table 5, the classification accuracy obtained using various feature selection methods with the top 10%, 15%, 20%, 25% and 30% selected features are given. For Page Block, Vertebral column, Ecoli, Breast cancer Wisconsin, Breast Tissue, seeds and Manual segmentation datasets, the number of features is small. After rounding the numbers features to be selected for top 10%, 15%, 20%, 25% and 30%, some of the numbers coincide, hence the result for these datasets contain less than 5 cases.

Classification accuracy was obtained using Naïve Bayes classifier with 10 fold cross validation. In each cross validation, feature selection methods are used to select the relevant features using only training data. Training is carried out using Naive Bayes on the selected features of the training data and accuracy of model is computed by using corresponding features of testing data. This procedure eliminates the selection bias [1] problem in selecting relevant features. The classification accuracy along with the standard deviation is given in Tables 6-13.

Table 5: Description of data sets

Data set	Total classes	Total features	Total patterns
Page Block	5	10	5473
Vertebral column	3	6	310
Ecoli	8	7	336
Ozone	2	72	1847
Gas sensor array drift	6	128	445
Steel Plate Faults	7	27	1941
Breast cancer Wisconsin	2	9	683
Manual Segmentation	4	15	16383

For each of the methods ReliefF, mRMR and chi2, t-test is performed under the null hypothesis that mean accuracy of the method is same as that of FSDD against the alternate hypothesis that mean accuracy of FSDD is greater than the corresponding method. The p-value of t-test is also given in the corresponding Tables 6-13 in parenthesis.

Table 6. Accuracy of methods for Page Block

Selected features	mRMR FCD	Chi square	ReliefF	FSDD
1	90.244 (0.397)	90.061 (0.244)	89.732 (0.027)	90.353
2	90.974 (0.217)	89.421 0.000)	90.389 (0.009)	91.449
3	91.631 ± (0.529)	87.650 ± (0.000)	90.115 ±(0.004)	91.595
Average	90.95	89.044	90.079	91.133

Table 7. Accuracy of methods for Vertebral column

Selected features	mRMR FCD	Chi square	ReliefF	FSDD
1	90.244 (0.39)	90.061 (0.24)	89.732 (.27)	90.353 (0.80)
2	90.974	89.421	90.389	91.449
3	91.631	87.650	90.115	91.595
Average	90.95	89.044	90.079	91.133

Table 8. Accuracy of methods for Ecoli dataset

Selected features	mRMR FCD	Chi square	ReliefF	FSDD
1	*64.743 (0.253)*	*66.848 (0.500)*	*66.848 (0.500)*	66.848
2	*64.695 (0.000)*	*64.725 (0.000)*	*64.725 (0.000)*	77.151
3	62.395 0.000)	77.427 (0.018)	77.164 (0.014)	82.360
Average	63.944	69.666	69.579	75.453

Table 9. Accuracy of methods for Ozone level detection

Selected features	mRMR FCD	Chi square	ReliefF	FSDD
8	67.456 (0.000)	67.348 (0.000)	66.218 (0.000)	75.960
11	68.160 (0.000)	66.103 (0.000)	67.511 (0.000)	76.499
15	67.454 (0.000)	64.153 (0.000)	66.322 (0.000)	73.846
18	65.938 (0.000)	64.529 (0.000)	66.863 (0.001)	71.625
22	67.238 (0.003)	65.396 (0.000)	67.347 2.469(0.002)	71.192
Average	67.249	65.506	66.852	73.824

Table 10. Accuracy of methods for Gas sensor array drift

Selected features	mRMR FCD	Chi square	ReliefF	FSDD
13	72.116 (0.377)	63.611 (0.000)	65.617 (0.001)	72.822
20	72.838 (0.406)	64.308 (0.000)	62.738 (0.000)	73.273
26	71.246 (0.450)	71.741 (0.555)	71.217 (0.451)	71.464
32	69.463 (0.233)	74.647 (0.968)	73.925 (0.937)	70.776
39	72.798 (0.499)	75.077 (0.845)	74.617 (0.804)	72.802
Average	71.692	69.877	69.623	72.227

Table 11. Accuracy of methods for Steel Plate Faults dataset

Selected features	mRMR FCD	Chi square	ReliefF	FSDD
1	89.745 (0.154)	87.412 (0.012)	89.901 (0.239)	90.911
2	94.729 (0.199)	95.172 (0.432)	94.722 (0.241)	95.310
3	96.340 (0.500)	95.315 (0.098)	96.046 (0.319)	96.340
Average	93.605	92.633	93.556	94.187

Table 12. Accuracy of methods for Breast cancer Wisconsin

Selected features	mRMR FCD	Chi square	ReliefF	FSDD
6	87.702 (0.091)	80.094 (0.000)	87.972 (0.170)	88.904
9	88.505 (0.151)	80.757 (0.000)	88.505 (0.155)	89.435
12	89.830 (0.500)	80.494 (0.000)	89.963 (0.542)	89.830
15	90.771 (0.584)	80.087 (0.000)	89.699 (0.260)	90.500
18	91.704 (0.707)	79.554 (0.000)	88.363 (0.049)	91.035
Average	89.702	80.197	88.901	89.941

Table 13. Accuracy of methods for Manual segmentation

Total selected features	mRMR FCD	Chi square	ReliefF	FSDD
2	87.302 (0.500)	86.949 (0.306)	86.949 (0.306)	87.302
3	87.055 (0.500)	86.503 (0.213)	87.055 (0.500)	87.055
4	86.805 (0.087)	86.805 (0.087)	86.754 (0.078)	87.704
5	86.807 (0.261)	86.857 (0.289)	86.904 (0.296)	87.259
Average	86.992	86.778	86.915	87.330

The p-values in Tables 6-13 show that for Ecoli and Steel Plate Faults datasets, performance of FSDD is better than other feature selection methods for various feature selection combinations. It can be asserted that for Ozone level detection dataset, FSDD performs better than other feature selection methods for all the selected features. For Breast cancer Wisconsin and Manual segmentation datasets FSDD performs at par with other methods for all the selected features. For Image segmentation dataset, performance of FSDD is better than that of mRMR and ReliefF for majority of selected features and better than chi2 for every combination of selected feature.

V Conclusions

In this paper, a new method for feature selection has been proposed. The method selects the features which can distinguish between two classes at a time. Instead of using some derived statistic, the density function of features has been used. Further, the relevant features have been chosen in conjunction with the inconsistency measure. The features, which are chosen as important by other feature selection methods, are also chosen by FSDD. In addition, FSDD also considers the features which play significant role in classification but are neglected by other feature selection methods. The classification accuracy of proposed method as compared to the existing feature selection methods, has been

shown to be better for majority of datasets taken from UCI machine learning repository and on manual segmentation data of NIST.

REFERENCES

[1] Ambroise, Christophe, and Geoffrey J. McLachlan. "Selection bias in gene extraction on the basis of microarray gene-expression data." Proceedings of the National Academy of Sciences 99.10 (2002): 6562-6566.

[2] Dash, Manoranjan, Huan Liu, and Hiroshi Motoda. "Consistency based feature selection." Knowledge Discovery and Data Mining. Current Issues and New Applications. Springer Berlin Heidelberg, 2000. 98-109.

[3] Ding et al. "Minimum redundancy feature selection from microarray gene expression data." Journal of bioinformatics and computational biology 3.02 (2005): 185-205.

[4] Ekenel et al. "Feature selection in the independent component subspace for face recognition." Pattern Recognition Letters 25.12 (2004): 1377-1388.

[5] Guyon et al. "An introduction to variable and feature selection." The Journal of Machine Learning Research 3 (2003): 1157-1182.

[6] Hall, Mark A., and Lloyd A. Smith. "Feature subset selection: a correlation based filter approach." (1997).

[7] Haury et al. "The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures." PloS one 6.12 (2011): e28210.

[8] Honavar et al. "Feature subset selection using a genetic algorithm." Intelligent Systems and their Applications, IEEE 13.2 (1998): 44-49.

[9] Liu, Huan, and Rudy Setiono. "Chi2: Feature selection and discretization of numeric attributes." Tools with Artificial Intelligence, 1995. Proceedings, Seventh International Conference on. IEEE, 1995.

[10] Marill et al. "On the effectiveness of receptors in recognition systems." Information Theory, IEEE Transactions on 9.1 (1963): 11-17.

[11] Mitra, P. ; Murthy, C. A. ; Pal, S. K. (2002) *Unsupervised feature selection using feature similarity* IEEE Transactions on Pattern Analysis and Machine Intelligence, 24 (3). pp. 301-312. ISSN 0162-8828

[12] Pantel et al. "Spamcop: A spam classification & organization program." Proceedings of AAAI-98 Workshop on Learning for Text Categorization. 1998.

[13] Robnik-Šikonja, Marko, and Igor Kononenko. "Theoretical and empirical analysis of ReliefF and RReliefF." Machine learning 53.1-2 (2003): 23-69.

[14] Serpico at al. "A new search algorithm for feature selection in hyperspectral remote sensing images." Geoscience and Remote Sensing, IEEE Transactions on 39.7 (2001): 1360-1367.

[15] Song, Le, et al. "Supervised feature selection via dependence estimation." Proceedings of the 24th international conference on Machine learning. ACM, 2007.

[16]Yang et al. "A comparative study on feature selection in text categorization." Machine Learning – International workshop- Morgan Kaufmann Publishers,inc.,1997.

[17] Wang, Yu, et al. "Gene selection from microarray data for cancer classification—a machine learning approach." Computational biology and chemistry 29.1 (2005): 37-46.

[18] Whitney et al. "A direct method of nonparametric measurement selection." Computers, IEEE Transactions on 100.9 (1971): 1100-1103.

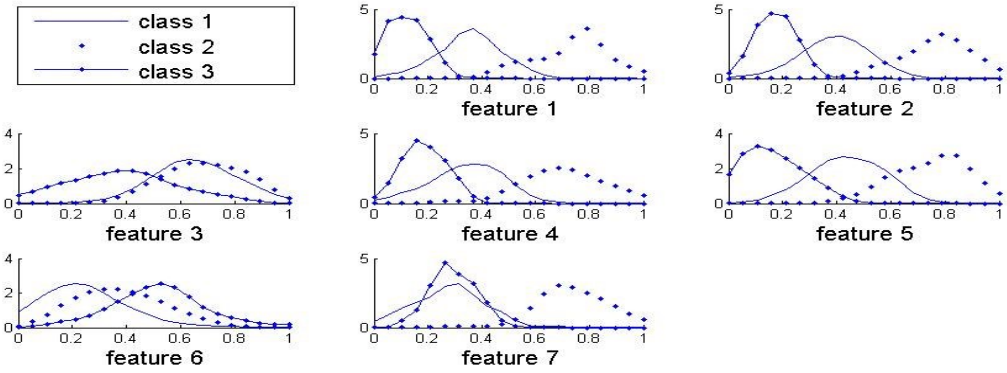

Fig.3 Density function of various features of Seeds dataset corresponding to the three classes

Table.4. 10 fold CV accuracy of various features set

mRMR		chi2		ReliefF		FSDD	
Features set	accuracy	Features set	accuracy	Features set	accuracy	Features set	accuracy
1	86.67	2	87.62	1	86.67	1	86.67
1, 2	86.19	2, 5	84.29	1, 2	86.19	1, 7	91.9
1, 2, 5	85.71	2, 5, 1	85.71	1, 2, 5	85.71	1, 7, 6	94.29
1, 2, 5, 7	89.05	2, 5, 1, 4	86.67	1, 2, 5, 7	89.05	1, 7, 6, 2	92.38
1, 2, 5, 7, 4	88.09	2, 5, 1, 4, 7	88.09	1, 2, 5, 7, 4	88.09	1, 7, 6, 2, 5	90.47