STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **9**, September 2021, pp 587–606.
Published online in International Academic Press (www.IAPress.org)

Feature Selection Based on Divergence Functions: A Comparative Classification Study

Saeid Pourmand [1], Ashkan Shabbak [2], Mojtaba Ganjali [1,*]

[1] *Department of Statistics, Faculty of Mathematical Sciences, Tehran, Iran*
[2] *Statistical Research and Training Center (SRTC), Tehran, Iran*

Abstract Due to the extensive use of high-dimensional data and its application in a wide range of scientific fields of research, dimensionality reduction has become a major part of the preprocessing step in machine learning. Feature selection is one procedure for reducing dimensionality. In this process, instead of using the whole set of features, a subset is selected to be used in the learning model. *Feature selection* (FS) methods are divided into three main categories: filters, wrappers, and embedded approaches. Filter methods only depend on the characteristics of the data, and do not rely on the learning model at hand. Divergence functions as measures of evaluating the differences between probability distribution functions can be used as filter methods of feature selection. In this paper, direct usages of several divergence functions such as *Jensen-Shannon* (JS) divergence and *Exponential divergence* (EXP) as FS methods are evaluated and performances of these divergence functions are compared with those of some of the most-known filter feature selection methods such as *Information Gain* (IG) and *Chi-Squared* (CHI). This comparison was made through accuracy rate and F1-score of classification models after implementing these feature selection methods.

Keywords Dimensionality Reduction, Machine Learning, Feature Selection, Filters, Wrappers, Embedded Methods, Divergence Functions.

AMS 2010 subject classifications 62H30, 62P99

DOI: 10.19139/soic-2310-5070-1092

1. Introduction

In recent years, dealing with high dimensional data has grown into a big part of machine learning and statistics including classification problems. This problem requires scientists to reduce the data dimension. One approach to dimension reduction is feature selection. Feature selection is a procedure in which a subset of the primary full set of features is selected to be applied in the learning process. The main goal of this selection procedure is to remove irrelevant and perhaps redundant features, but it also helps improve learning performance, prevents overfitting, and also reduces the computational costs [20]. There are multiple ways to perform FS, but in general, this procedure is classified into three main categories [11]: filters, wrappers, and embedded methods. Filter methods only employ data to perform FS. This group of methods has no dependencies on the classifier and hence has lower time complexity than other groups of methods. IG [27] and CHI [23] are two of the most widely used filter methods. Wrapper methods employ the classifier to evaluate features and reach the potentially suitable subset of features. These methods perform classification on a different subset of features, then for each subset assess the performance of the classifier and select the subset which resulted in the best performance. Consequently, having higher time complexity than filter methods is inevitable for wrapper methods. Moreover, they usually tend to result in a better

[*] Correspondence to: Dr. Mojtaba Ganjali, Professor of Statistics, (Email: m-ganjali@sbu.ac.ir). Department of Statistics, Faculty of Mathematical Sciences. Shahid Beheshti University, Tehran, Iran. (Tel: 0098-21-29902915).

ISSN 2310-5070 (online) ISSN 2311-004X (print)

outcome since they examine all possible subsets of features. Embedded methods [18] operate differently from antecedent methods. These methods are placed inside the classifier and are merged with the learning process.

The main advantages of filter methods over others are their low time complexity, compared with the other two categories of methods, and their resistance to overfitting [10]. The reason behind these differences is that filter methods are part of a preprocessing step to the learning problem while the other groups of methods are deployed alongside or within the learning process.

Most FS techniques are established based on measuring the difference between observed and expected probabilities. For instance, two popular methods, which this statement holds for, are IG and CHI. In this paper, we review some FS methods that are compatible with this description as well as IG and CHI. At first, we have to provide some details about expected and observed probabilities. Without loss of generality, throughout this paper, X and Y are considered categorical variables (In cases where X is continuous, it can be categorized using the equal-width binning method). Thus the learning framework is a classification problem. $P(X)$ and $P(Y)$, are true marginal distributions of X and Y respectively, and $P(X, Y)$ is the joint probability distribution function of X and Y. In real-world problems, these probability distributions are unknown and have to be estimated using observations. Suppose that $p(x)$, $p(y)$, and $p(x, y)$ are estimated probabilities. We have:

$$p(x) = \frac{|x|}{N}, \qquad p(y) = \frac{|y|}{N}, \qquad p(x, y) = \frac{|x, y|}{N},$$

where $|x|$ is the number of times the variable X is equal to x. Definitions are similar for $|y|$ and $|x, y|$. And N is the size of observations in data.

When X and Y are independent, the joint probability will equal $P(X)P(Y)$, so the difference between two probabilities $P(X, Y)$ and $P(X)P(Y)$ provide useful interpretation about the relationship between X and Y. One way to evaluate this difference is to use divergence functions. In statistics, divergence functions are used to assess the distance or difference between two probability distribution functions. Divergence functions for each probability distribution P and Q from the space of probability distributions S must satisfy [5]:

- $D(P, Q) \geq 0, \quad \forall P, Q \in S,$
- $D(P, Q) = 0, \quad if\ P = Q,$

where $D(P, Q)$ is the divergence function of two probability distribution functions P and Q. Note that divergences are not necessarily metrics, but under some conditions, metrics defined over probability spaces can be divergence functions. IG and CHI are two instances of FS based on divergence functions. In IG, the amount of information gained from target variable Y given feature X can be formulated as below:

$$IG(Y, X) = \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)}, \tag{1}$$

where N_X is the number of categories in feature X and N_Y is the number of classes in target variable Y.

Evidently, it can be seen from Eq.(1) that if the joint probability distribution $p(x, y)$ and product of marginal distributions of X and Y i.e. $p(x)p(y)$ were close to each other, $IG(Y, X)$ would get close to zero. This means that we gain little information about Y provided that X is observed.

The right-hand side of Eq.(1) is obtained from the Kullback-Leibler (KL, [16]) divergence function which is defined as below:

$$KL(X, Y) = \int\int P(x, y) \log \frac{P(x, y)}{Q(x, y)} \mathrm{d}x\mathrm{d}y. \tag{2}$$

CHI FS method is based on the chi-squared test of independence and is defined as follows:

$$CHI(X, Y) = \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} \frac{[p(x_i, y_j) - p(x_i)p(y_j)]^2}{p(x_i)p(y_j)}. \tag{3}$$

Similar to Eq.(1), in Eq.(3) if $p(x_i, y_j)$ and $p(x_i)p(y_j)$ for all i and j were close to each other, $CHI(X, Y)$ tends to zero. This method is based on another divergence function called Kagan's Divergence [26] that can be formulated as below:

$$\mathcal{X}^2(X,Y) = \int\int \frac{(P(x,y) - Q(x,y))^2}{Q(x,y)} \mathrm{d}x\mathrm{d}y. \tag{4}$$

1.1. Related works

A lot of work has been done about using divergence functions as a tool to improve learning performance. Schneider [28], Wang et al. [32], Zhang et al. [34], and Lee et al. [19] used the Kullback-Leibler divergence function to improve FS performance. Lifang et al. [21] proposed a hierarchical class correlation feature selection method based on Kullback-Leibler divergence. Jiang et al. [15] Developed a hybrid feature selection method based on Kullback-Leibler divergence methods. Fu et al. [8] used Hellinger distance to achieve stable sparse feature selection applied to class-imbalanced data. Hart et al [13] used Hellinger distance to rank features in a genetic programming framework. Hashemi et al. [14] used L2-distance to rank features in multi-label datasets. Temrat et al. [30] performed feature selection based on total variation distance for OSA classification. Yoon et al. [33] applied an estimation of Jensen-Shannon divergence to capture locally important features. Cui et al. [6] Proposed a new information-theoretic criterion based on Jensen-Shannon divergence measure between the probability distributions of the random walk on different graphs of features. Novovicov et al. [25] used the Kullback-J divergence to build an FS method for multi-modal binary classification problems. Guzmn-Martnez and Alaiz-Rodrguez [12] used the Jensen-Shannon divergence to assess the stability of FS methods. Niazi et al. [24] used the Kullback-Leibler divergence-based FS method for a power system security classification problem. Kumar and Minz [17] briefly reviewed divergence functions as FS methods. Thabtah et al. [29] proposed the L2 divergence function as an FS method and compared its performance with those of IG and CHI FS methods.

2. Divergence functions

Divergence functions in statistics can be used to measure the difference between two probability functions. A divergence function takes two probability distribution functions as inputs and results in a non-negative real-valued output.

In a learning problem, these two probability distribution functions (PDFs) can be expected PDF and observed PDF, which in this paper are indicated as P and Q, respectively. In an FS framework P is $p(x,y)$ and Q is $p(x)p(y)$. Therefore the output of a divergence function for a feature X and target variable Y, compared to the result for other features, determines if the feature should be selected or not. In other words, if these two PDFs are close to each other for feature X and target variable Y, it indicates that this feature has little effect on the target variable and perhaps should be discarded.

Divergence functions are divided into three main groups that will be explained briefly.

2.1. f-divergences

In general, for two $P(x)$ and $Q(x)$, an f-divergence function can be defined as [2]:

$$I_f(P,Q) = \int Q(x) f(\frac{P(x)}{Q(x)}) \mathrm{d}x, \tag{5}$$

in which f must have these properties:

- f must be convex,
- $f(1) = 0,\ f'(1) = 0,\ f''(1) = 1.$

Kullback-Leibler divergence, Hellinger distance [31], Kagan's divergence, and Jensen-Shannon divergence [9] are some of the most known f-divergences.

2.2. Bregman's divergences

This family of divergence functions was generally introduced by Bregman in [3]. As [2] stated, for each two probability distribution functions $P(x)$ and $Q(x)$, divergence functions in this family are defined as follows:

$$D_{\varphi}(P,Q) = \int [\varphi(P(x)) - \varphi(Q(x)) - P(x)Q(x)\varphi'(Q(x))]\mathrm{d}x, \tag{6}$$

such that φ is a differentiable strictly convex function.

2.3. α-divergences

α-divergences are another group of divergence functions that are parameterized on $\alpha \in (-\infty, +\infty)$. The basic asymmetric α-divergence function can be defined as [5]:

$$D^{\alpha}(P,Q) = \frac{1}{\alpha(\alpha-1)} \int [P^{\alpha}(x)Q^{1-\alpha}(x) - \alpha P(x) + (\alpha-1)Q(x)]\mathrm{d}x, \tag{7}$$

where $\alpha \in \mathbb{R} \setminus \{0,1\}$ and P and Q do not need to be normalized. These divergence functions can be derived from both f-divergences and Bregman's divergences [1]. One of the most-known α-divergence functions is Rényi divergence [31].

2.4. Implemented divergence functions for feature selection

In this paper, we compare some of the prominent divergence functions in the context of FS. The classification of these divergence functions and their corresponding formulas in FS are available in Table 1.

3. Experimental Design

In this section, we discuss the datasets, how divergence-based FS methods work, and the general classification framework used in this paper.

The datasets are all gathered from the UCI machine learning repository [7]. Table 2 provides some useful information about these datasets. This information includes the name of the datasets, the number of inheritances they used, the number of features they include, the number of classes their response of interest has, whether they have missing values or not, and whether they include any continuous variable or not.

Divergence-based FS methods rank features based on their divergence measure. The problem with ranker FS methods is that they do not provide a subset of features, as they only rank them based on a measure. Therefore, a threshold is needed to select a suitable subset of features.

Thresholds can be selected from percentiles of the feature divergence measures. As an example, it can be selected from the set $\{5\%, 10\%, \cdots, 95\%, 100\%\}$. In this set, 5% indicates selecting the top 5% of features ordered based on a predetermined and calculated divergence measure, and 100% indicates all of the features. Other percentiles in this set are defined similarly. Different Percentiles result in different thresholds and a subsequently different subset of features. Therefore, these percentiles must be chosen wisely. One way is to choose the thresholds empirically based on the classification results on some datasets and then use them in general.

Divergence	General formula	FS formula
Kullback-Leibler Divergence	$KL(X,Y) = \int P(x) \log \frac{P(x)}{Q(x)} \mathrm{d}x$	$\sum_{i,j} p(x_i, y_j) \log \frac{p(x_i,y_j)}{p(x_i)p(y_j)}$
Kagan's Divergence	$\mathcal{X}^2(X,Y) = \int \frac{(P(x)-Q(x))^2}{Q(x)} \mathrm{d}x$	$\sum_{i,j} \frac{[p(x_i,y_j)-p(x_i)p(y_j)]^2}{p(x_i)p(y_j)}$
Hellinger Distance	$HE(X,Y) = \int (\sqrt{P(x)} - \sqrt{Q(x)})^2 \mathrm{d}x$	$\sum_{i,j} = \sqrt{p(x_i, y_j)} - \sqrt{p(x_i)p(y_j)})^2$
Jensen-Shannon Divergence	$\frac{1}{2}KL_{pm}(X,Y) + \frac{1}{2}KL_{qm}(X,Y)$	$\frac{1}{2}KL_{pm}(X,Y) + \frac{1}{2}KL_{qm}(X,Y)$
L2 Divergence [29]	$L^2(X,Y) = \int (P(x) - Q(x))^2 \mathrm{d}x$	$\sum_{i,j} [p(x_i, y_j) - p(x_i)p(y_j)]^2$
Total Variation Divergence [26]	$TV(X,Y) = \int \lvert P(x) - Q(x) \rvert \mathrm{d}x$	$\sum_{i,j} \lvert p(x_i, y_j) - p(x_i)p(y_j) \rvert$
Exponential Divergence [4]	$EXP(X,Y) = \int P(x)[\ln(P(x)) - \ln(Q(x))]^2$	$\sum_{i,j} p(x_i, y_j)[\ln(p(x_i, y_j)) - \ln(p(x_i)p(y_j))]^2$

Table 1. Overall definition of divergence functions and their FS form

Data	**# of instances**	**# of features**	**# of classes**	**Missing values**	**Continuous variables**
Arrhythmia	452	279	16	No	Yes
Audiology	200	69	24	Yes	No
Autos	205	25	7	Yes	Yes
Bands	540	39	2	Yes	Yes
Car	1728	6	4	No	No
Cleveland	303	13	2	Yes	No
Colic	368	27	2	Yes	Yes
Credit-a	690	14	2	Yes	Yes
Dermatology	366	34	6	Yes	Yes
Flags	194	29	8	No	Yes
Glass	214	9	7	No	Yes
Hepatitis	155	19	2	Yes	Yes
Ionosphere	351	34	2	No	Yes
Lymphography	148	18	4	No	Yes
Mushroom	8124	22	2	Yes	No
Nursery	12960	8	5	No	No
Optic	5620	64	10	No	Yes
Page-Blocks	5473	10	5	No	Yes
Segment	2310	19	7	No	Yes
Sonar	208	60	2	N/A	Yes
Vote	435	16	2	Yes	No

Table 2. The datasets description

To show how this is proceeded, first we need to describe the classification framework. After the FS process, a classifier performs the classification task. here, to select the suitable threshold for FS methods, the C5.0 decision tree classifier is used along with a 10-fold cross-validation technique. Later, the naïve Bayes classifier is used in the same situation for all datasets to evaluate the effect of these FS methods on each of the datasets.

There are many metrics available to evaluate the performance of a classifier: accuracy, F1-Score, precision, and recall are just a few of them. These metrics can be obtained and calculated from a confusion matrix. A confusion matrix in a binary classification problem consists of four elements: TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative) which indicate the number of positive cases that are predicted positive, negative cases that are predicted negative, negative cases that are predicted positive and positive cases that are predicted negative, respectively.

Each of the above-mentioned metrics can be calculated based on the elements of a confusion matrix:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}, \tag{8}$$

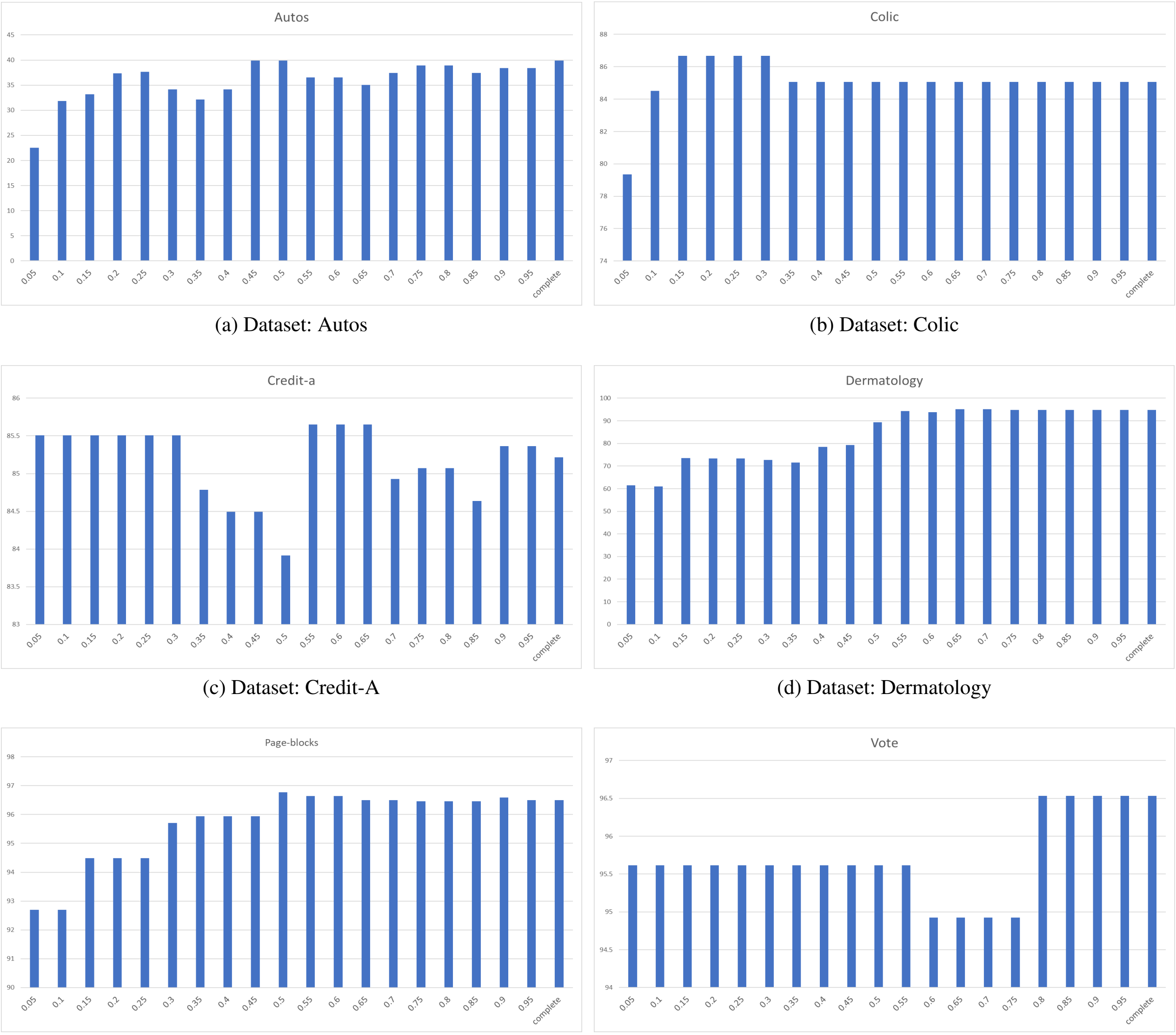
$$\text{Precision} = \frac{TP}{TP+FP}, \tag{9}$$

$$\text{Recall} = \frac{TP}{TP+FN}, \tag{10}$$

$$\text{F1-Score} = \frac{2 * Precision * Recall}{Precision + Recall}. \tag{11}$$

The calculations are similar in a multi-class problem. Therefore, accuracy as one of the aforementioned metrics is used to determine a suitable threshold for each FS method. Figure 1 displays how different threshold choices result in different classification accuracy on a few datasets. It should be noted that this demonstration is only attributable to presentation purposes. Also, to deal with continuous features, equal-width binning was used. Figures 1-7 show the classification accuracy results for each method on 6 different datasets. Figure 1 shows the classification accuracy of naïve Bayes classifier after performing the KL FS method with different percentiles as a chosen threshold on 6 different datasets. As can be seen, the results are stable by the 50% threshold most of the time. There is a decrease and instability by 50% in two datasets “Credit-A” and “Vote”, where in both of those, the amount of difference in accuracy is too small. With cutpoints less than 50%, the results become more unstable, as the amount of difference in accuracy becomes large. Therefore, 75% and 50%, which indicate choosing the top 75% and 50% features of the ranked set respectively, are the best choices among the set of percentiles. This choice is similar for other FS methods.

Choosing the same percentile as a threshold for all methods leads to selecting the same amount of features in most datasets. This means that, although they may select different subsets of features, they selected them with the same size.



(a) Dataset: Autos

(b) Dataset: Colic

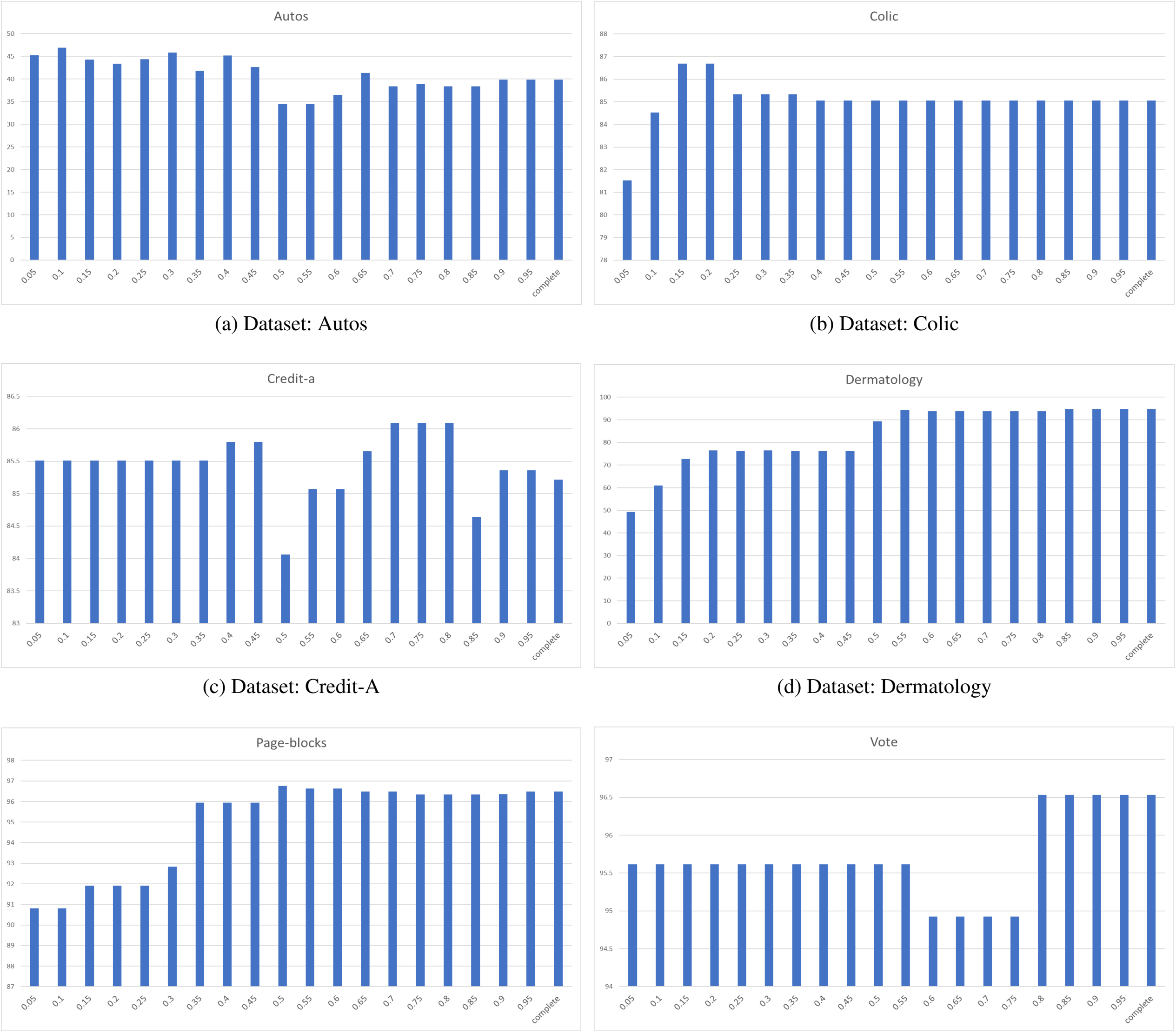
(c) Dataset: Credit-A

(d) Dataset: Dermatology

(e) Dataset: Page-Blocks

(f) Dataset: Vote

Figure 1. Classification accuracy percentage of datasets Autos, Colic, Credit-a, Dermatology, Page-Blocks and Vote used to induce suitable thresholds: KL method



(a) Dataset: Autos

(b) Dataset: Colic

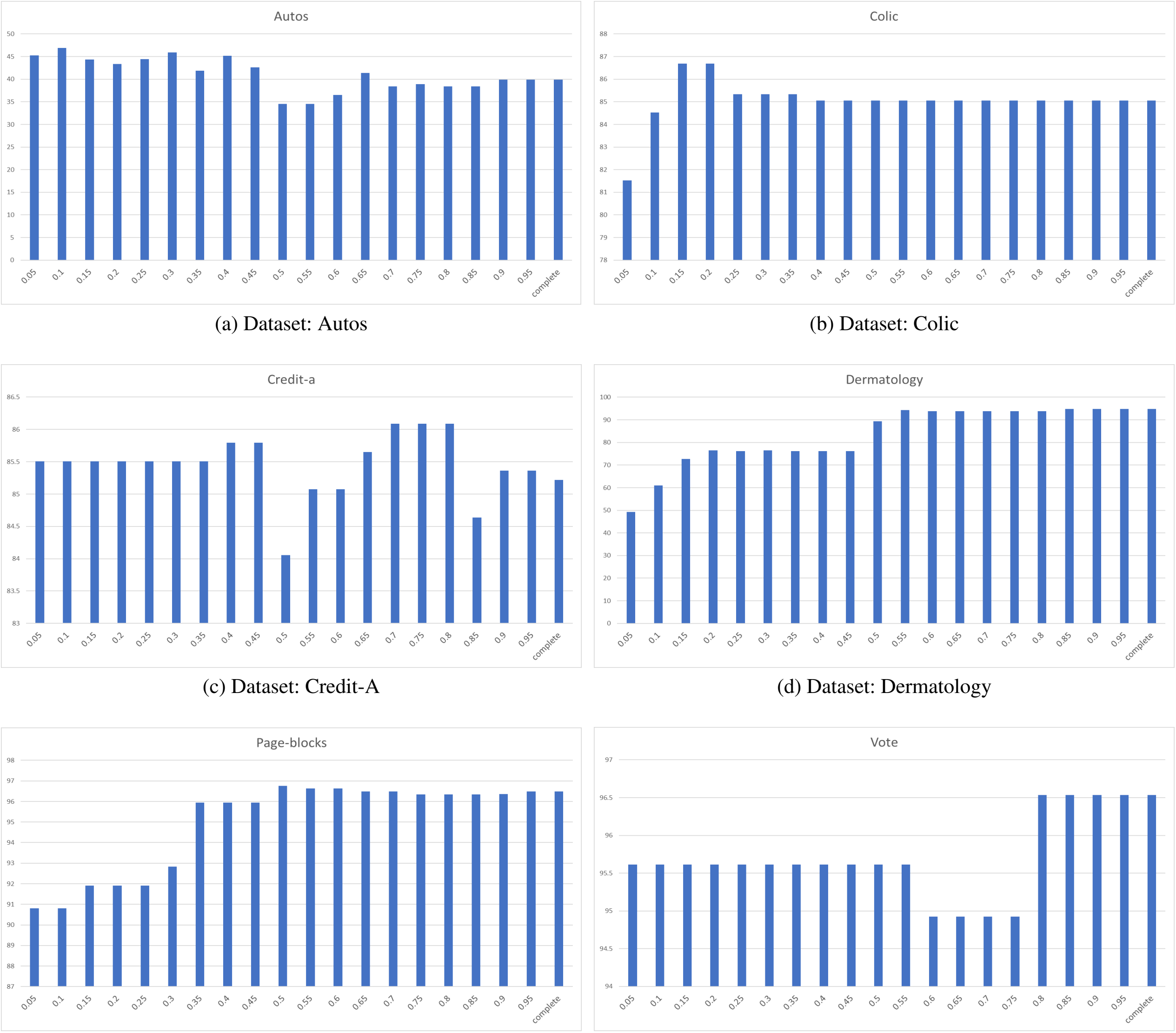
(c) Dataset: Credit-A

(d) Dataset: Dermatology

(e) Dataset: Page-Blocks

(f) Dataset: Vote

Figure 2. Classification accuracy percentage of datasets Autos, Colic, Credit-a, Dermatology, Page-Blocks and Vote used to induce suitable thresholds: L2 method



(a) Dataset: Autos

(b) Dataset: Colic

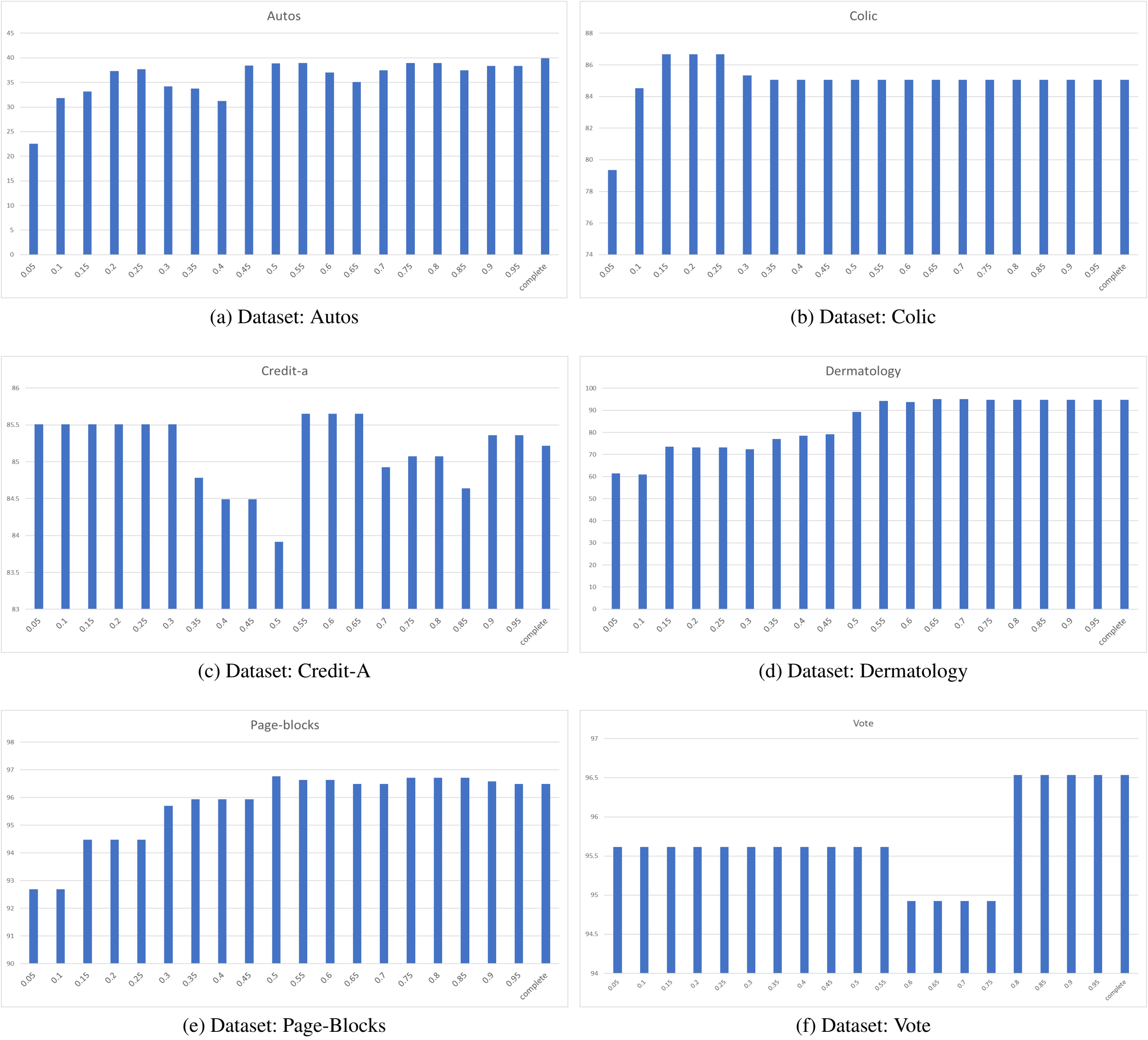
(c) Dataset: Credit-A

(d) Dataset: Dermatology

(e) Dataset: Page-Blocks

(f) Dataset: Vote

Figure 3. Classification accuracy percentage of datasets Autos, Colic, Credit-a, Dermatology, Page-Blocks and Vote used to induce suitable thresholds: CHI2 method



(a) Dataset: Autos

(b) Dataset: Colic

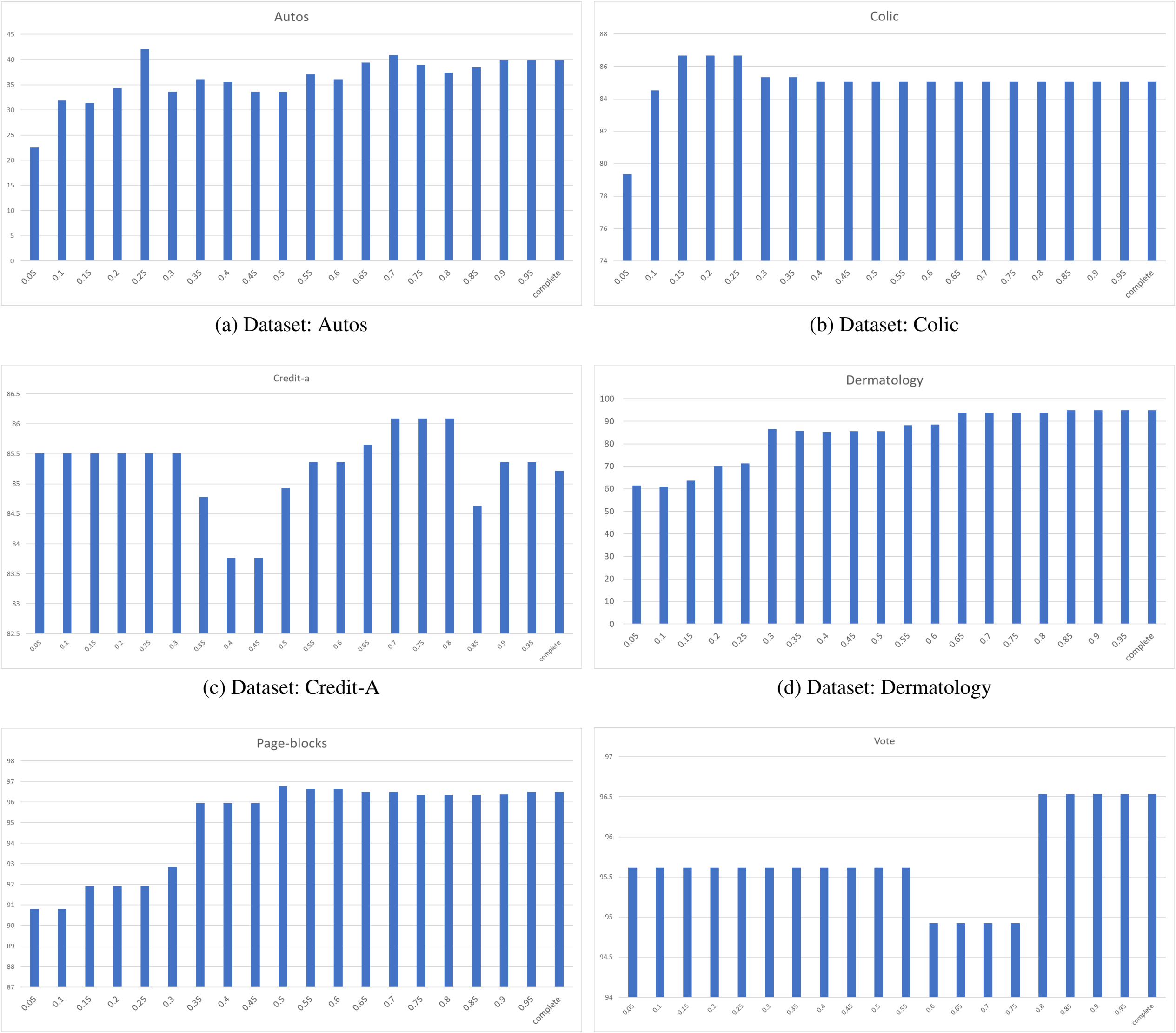
(c) Dataset: Credit-A

(d) Dataset: Dermatology

(e) Dataset: Page-Blocks

(f) Dataset: Vote

Figure 4. Classification accuracy percentage of datasets Autos, Colic, Credit-a, Dermatology, Page-Blocks and Vote used to induce suitable thresholds: HL method



(a) Dataset: Autos

(b) Dataset: Colic

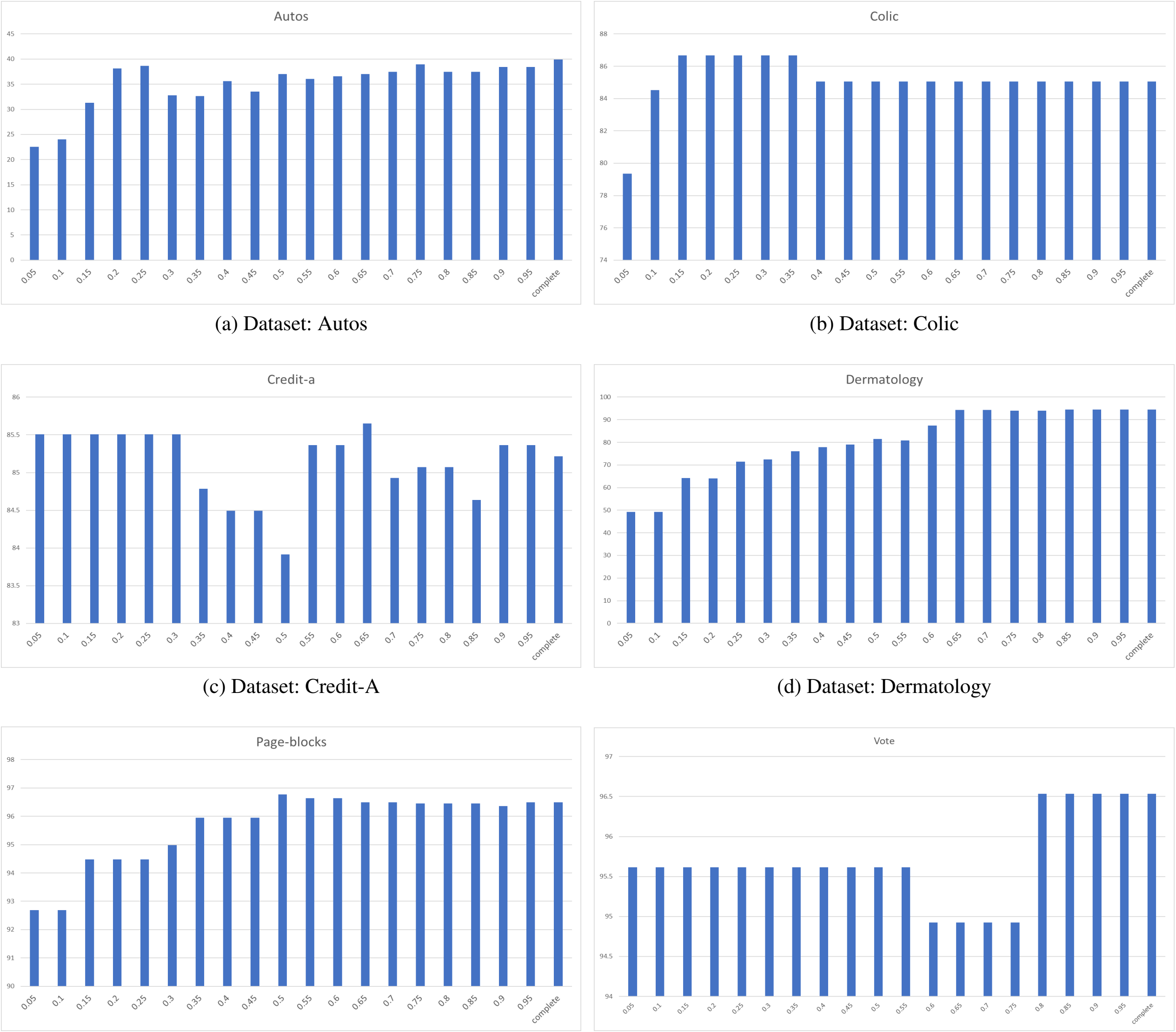
(c) Dataset: Credit-A

(d) Dataset: Dermatology

(e) Dataset: Page-Blocks

(f) Dataset: Vote

Figure 5. Classification accuracy percentage of datasets Autos, Colic, Credit-a, Dermatology, Page-Blocks and Vote used to induce suitable thresholds: TV method



(a) Dataset: Autos

(b) Dataset: Colic

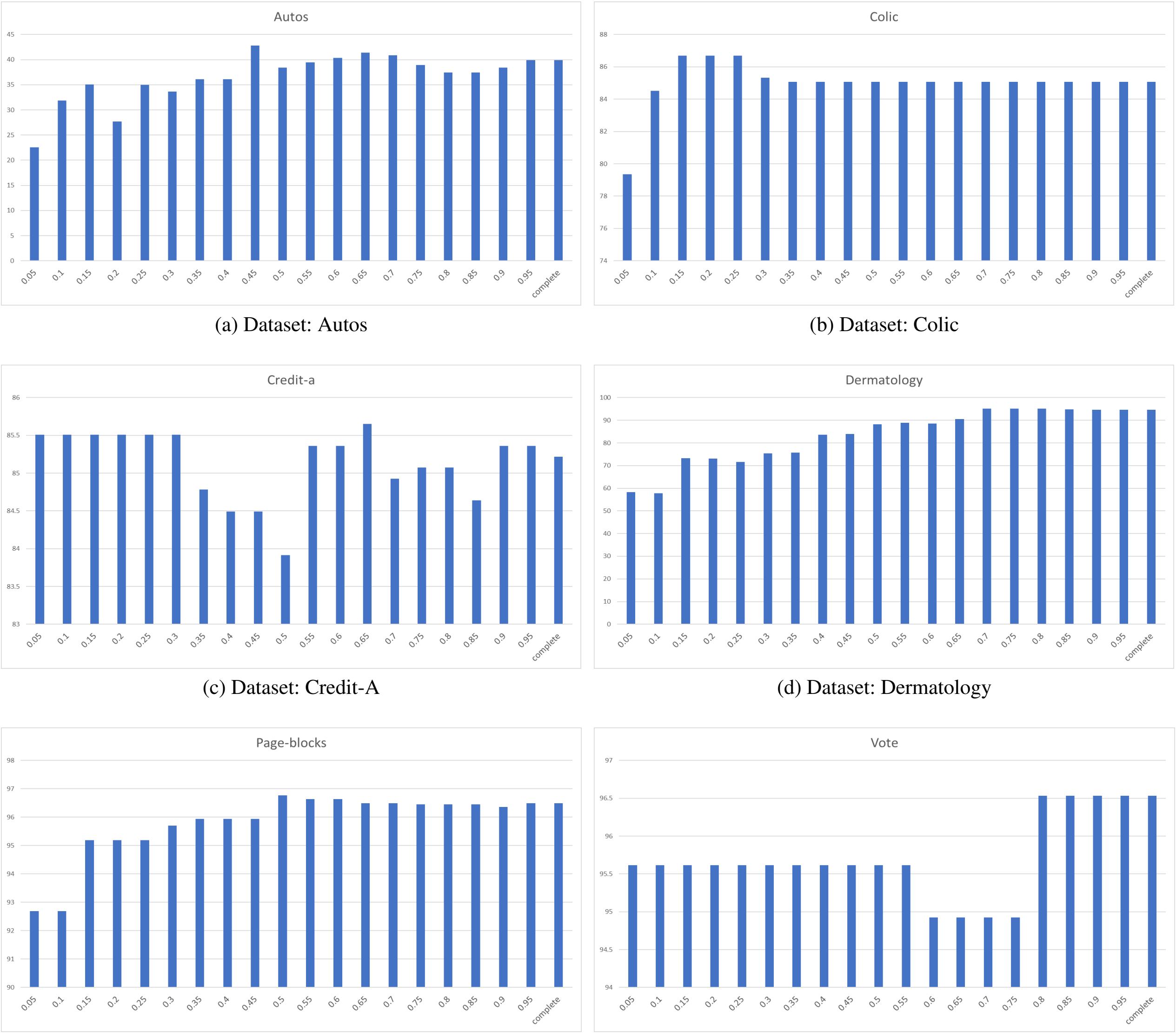
(c) Dataset: Credit-A

(d) Dataset: Dermatology

(e) Dataset: Page-Blocks

(f) Dataset: Vote

Figure 6. Classification accuracy percentage of datasets Autos, Colic, Credit-a, Dermatology, Page-Blocks and Vote used to induce suitable thresholds: EXP method



(a) Dataset: Autos

(b) Dataset: Colic

(c) Dataset: Credit-A

(d) Dataset: Dermatology

(e) Dataset: Page-Blocks

(f) Dataset: Vote

Figure 7. Classification accuracy percentage of datasets Autos, Colic, Credit-a, Dermatology, Page-Blocks and Vote used to induce suitable thresholds: JS method

This empirical demonstration is one possible way to identify and determine the potentially suitable threshold points. In that case, for all of the datasets above two threshold points look more promising than the others, 50%, and 75% thresholds. These cutoff points select the top 50% and 75% of the decreasingly-ordered feature sets, respectively. As it can be seen from this limited example, beginning and ending points in the threshold set do not result in a consistently good classification accuracy, such as the results from “Autos” and “Page-Blocks” datasets. Therefore, in further experiments, we used these two cutoff points as potential thresholds for all FS methods.

The naïve Bayes classifier is the classification method used in this paper from here on. This method is performed along with a 10-fold cross-validation technique. Also, to deal with continuous variables in the FS process, all continuous variables were converted to discrete variables using equal-width binning.

4. Results Analysis

4.1. Average number of selected features

All divergence-based FS methods mentioned in Table 1 were performed on datasets available in Table 2. And as the result, the average number of selected features for each method and each threshold over all datasets are available in Figure 8. Based on Figures (1-7) and Figure 8, it can be inferred that although there is a diversity in selected features based on each FS method, there is a similarity in the number of features selected most of the time. As can be seen, only three methods in 75% and two methods in 50% resulted in a slightly higher average number, which arises due to the number of features in two datasets “Arrhythmia” and “Audiology”. With 75% threshold and in “Arrhythmia” all methods selected 209 features except for CHI2 and HL which selected 210 and 211 features, respectively. Moreover, in “Audiology”, all methods selected 52 features besides the TV method. Furthermore, with a 50% threshold, in “Arrhythmia” all methods select 52 features besides JS and in “Audiology” all methods select 35 features except CHI2 which selects 36 features.

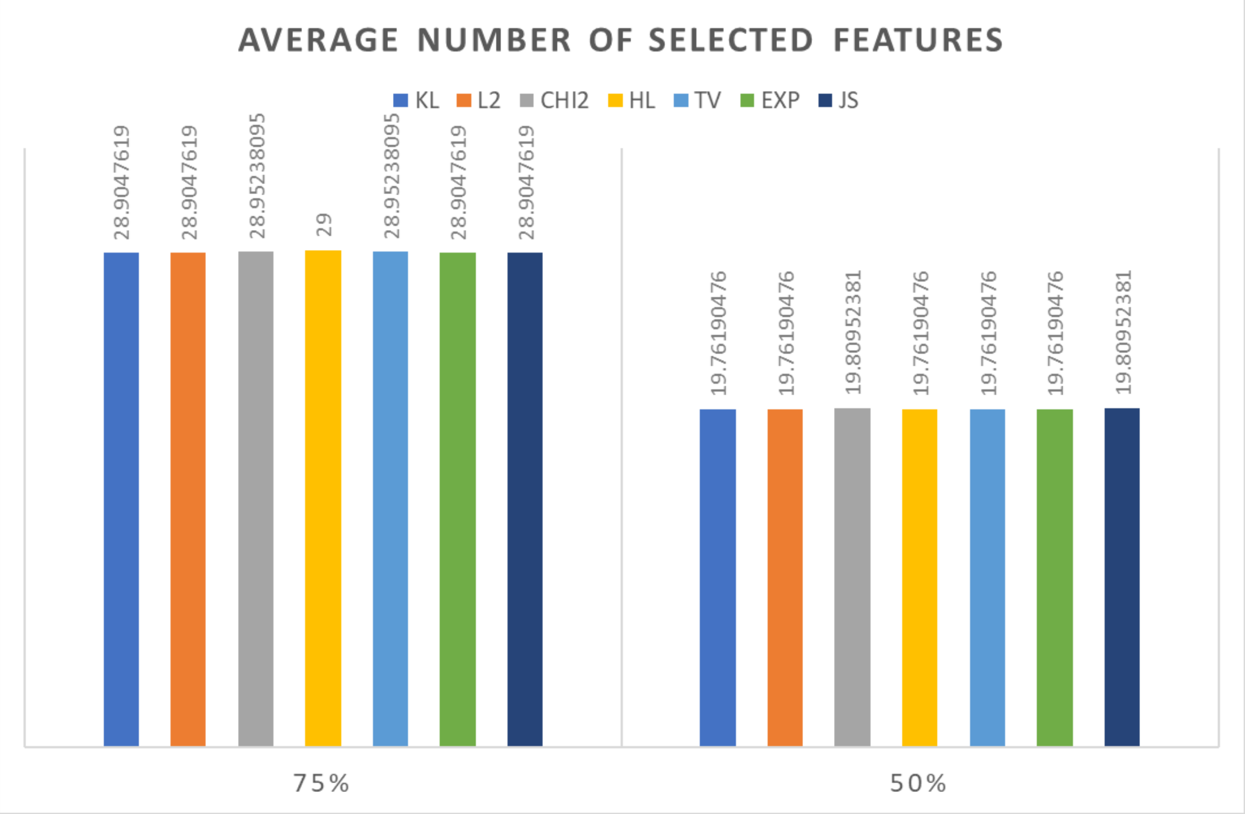
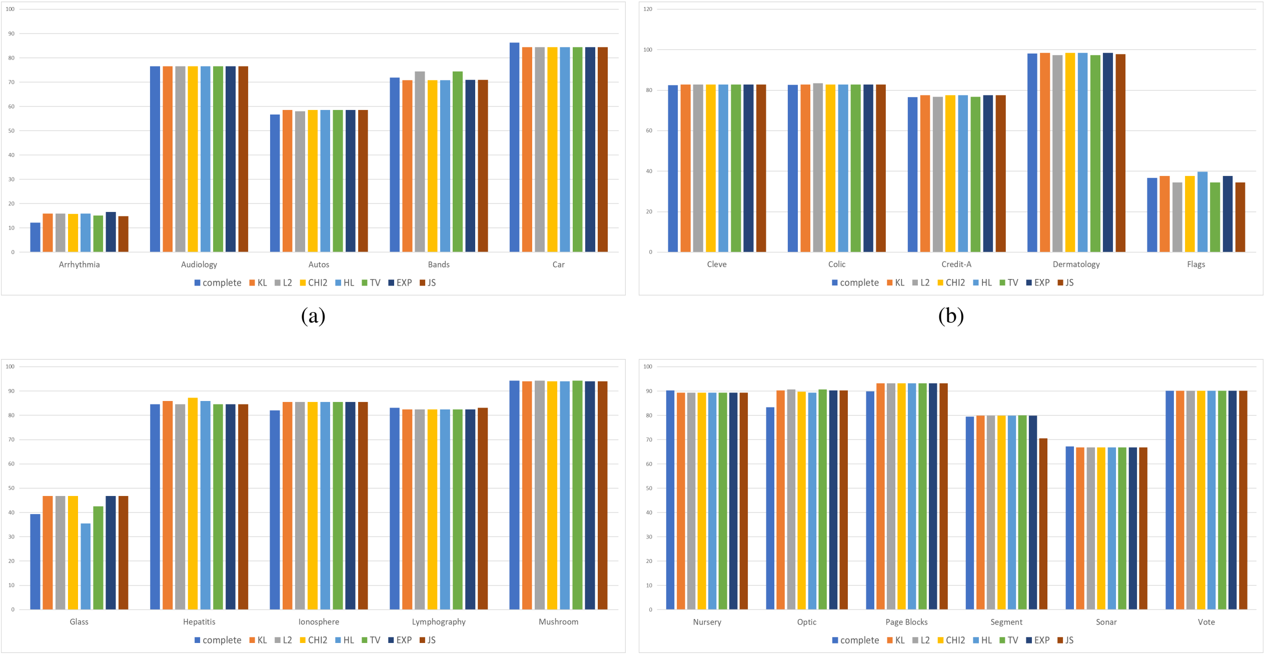


Figure 8. Average number of selected features for each divergence-based FS method

4.2. Accuracy

Figure 9 shows the classification accuracy results after performing FS methods. The selection procedure starts with ranking features based on each divergence-based method, then selecting the top 75% of ranked features. EXP divergence function led to the best results among all FS methods. The win-lose-tie record of EXP against other methods were 3-1-17, 5-5-11, 3-1-17, 5-2-14, 5-4-12, and 5-2-14 against KL, L2, CHI2, HL, TV, and JS, respectively. After that, L2 holds the second spot. Win-lose-tie records of this method were 5-5-11 against KL, 6-5-10 against CHI2, 6-5-10 against HL, 5-2-14 against TV, 5-5-11 against EXP, and 6-5-10 against JS. After L2, CHI2, KL, HL, TV, and JS are ranked respectively. Furthermore, KL and CHI recorded 13-6-2 against the complete set of features which was the best record. After these methods, HL, and EXP recorded 12-7-2, then TV and L2 recorded 11-8-2 and at last, the JS algorithm led to the worst record against the complete set of features among all methods with 10-8-3. On an average of 21 datasets, each FS method selects near 29 features. HL selects more features than other methods on average with exactly 29 features. Each method increases classification accuracy by around 1%, which means that with 75% of features these methods manage to increase the classification accuracy by 1%. CHI2, KL, and EXP led to the highest classification accuracy increase on average by 1.28%, 1.25%, and 1.23% respectively. The least increase in classification accuracy was led by the JS algorithm.



(a)

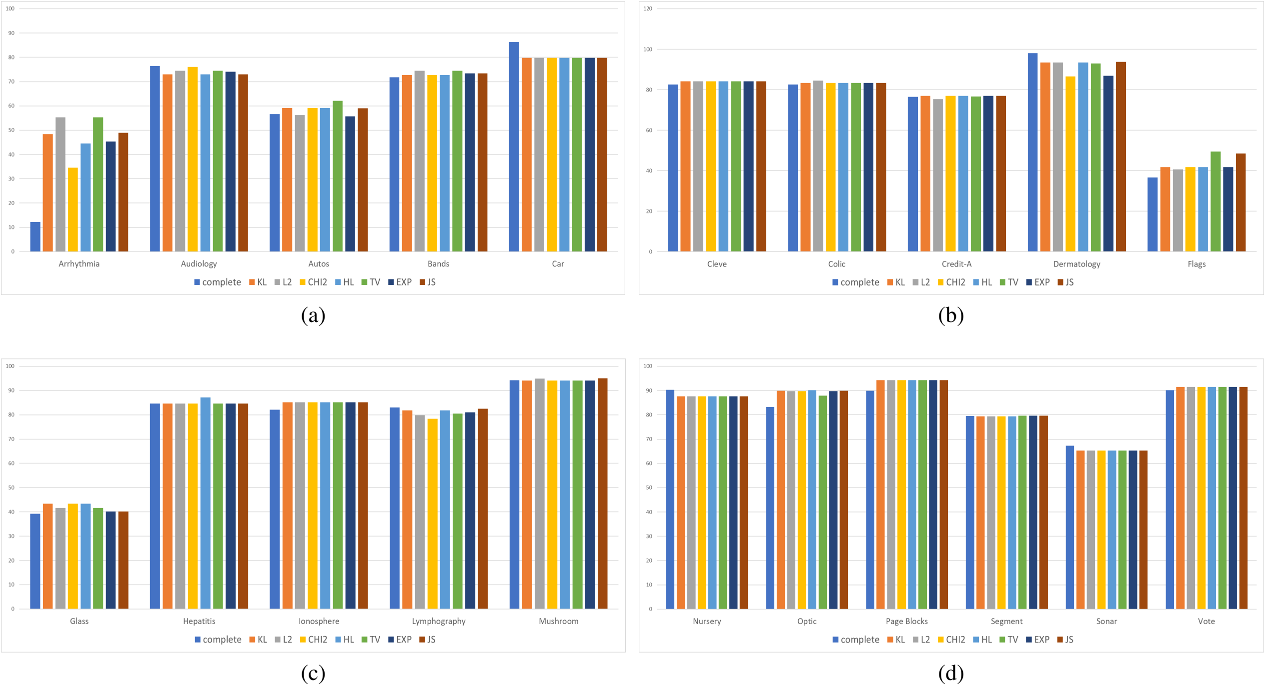
(b)

(c)

(d)

Figure 9. Classification accuracy percentage after performing all FS methods on each dataset and a 75% threshold

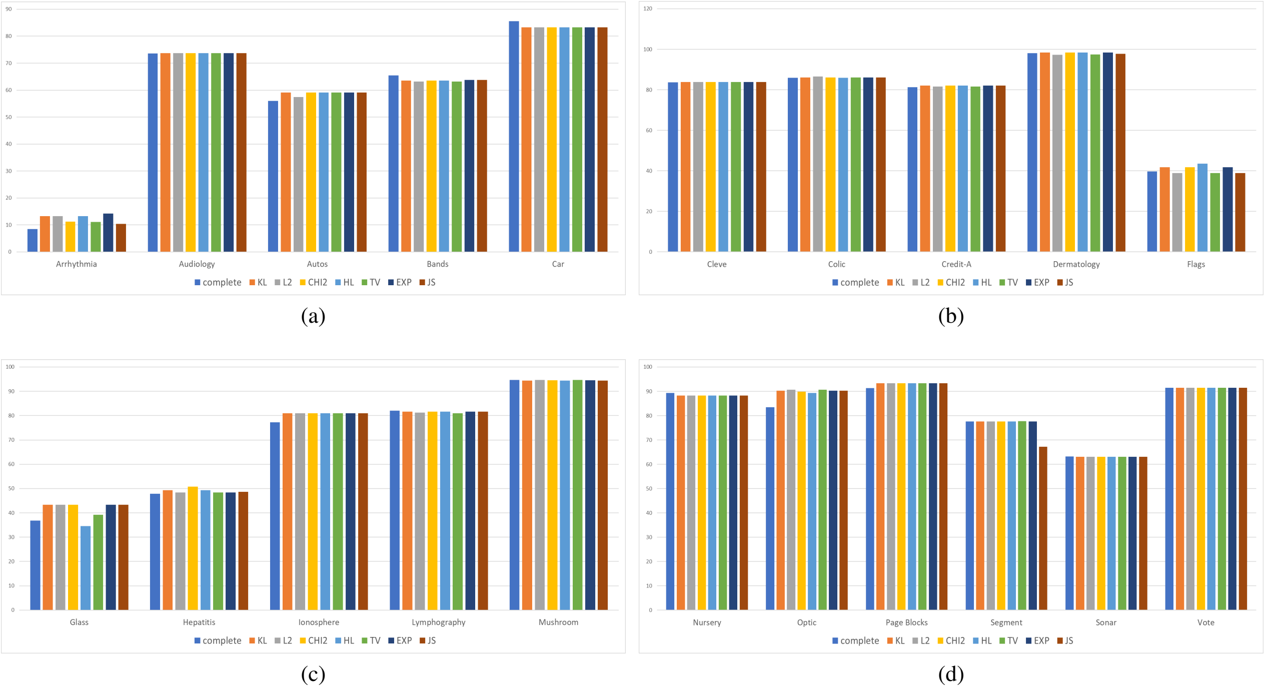
Similar to Figure 9, Figure 10 shows the classification results after performing divergence-based FS methods and selecting the top 50% of ranked features based on each method. Therefore a naïve Bayes was performed and its accuracy was calculated. Among all methods, the best classification performance occurred in the JS FS method. Win-lose-tie record of JS was 8-2-11, 8-5-8, 8-3-10, 9-4-8, 6-6-9, and 7-1-13 against KL, L2, CHI2, HL, TV, and EXP, respectively. This method also recorded 15-6-0 against the case of no FS which was better than all other methods. Furthermore, JS increased classification accuracy by approximately 2.54% on average, which was only behind the TV with a 2.94% increase. After JS, it was the TV that led to results better than all remaining methods, with records of 6-5-10, 5-5-11, 7-5-9, 7-6-8, 7-4-10, and 6-6-9 against KL, L2, CHI2, HL, EXP, and JS, respectively. This method resulted in better classification accuracies than a complete set of features in 14 out of 21 datasets, which was only behind results obtained by JS. Also, this method increased the classification accuracy by approximately 2.94% on average, which was higher than all other methods. After TV, L2, KL, HL, EXP, and CHI2, CHI2 had the lowest increase in accuracy on average with approximately 1.24%.



(a) (b)

(c) (d)

Figure 10. Classification accuracy percentage after performing all FS methods on each dataset and a 50% threshold



(a) (b)

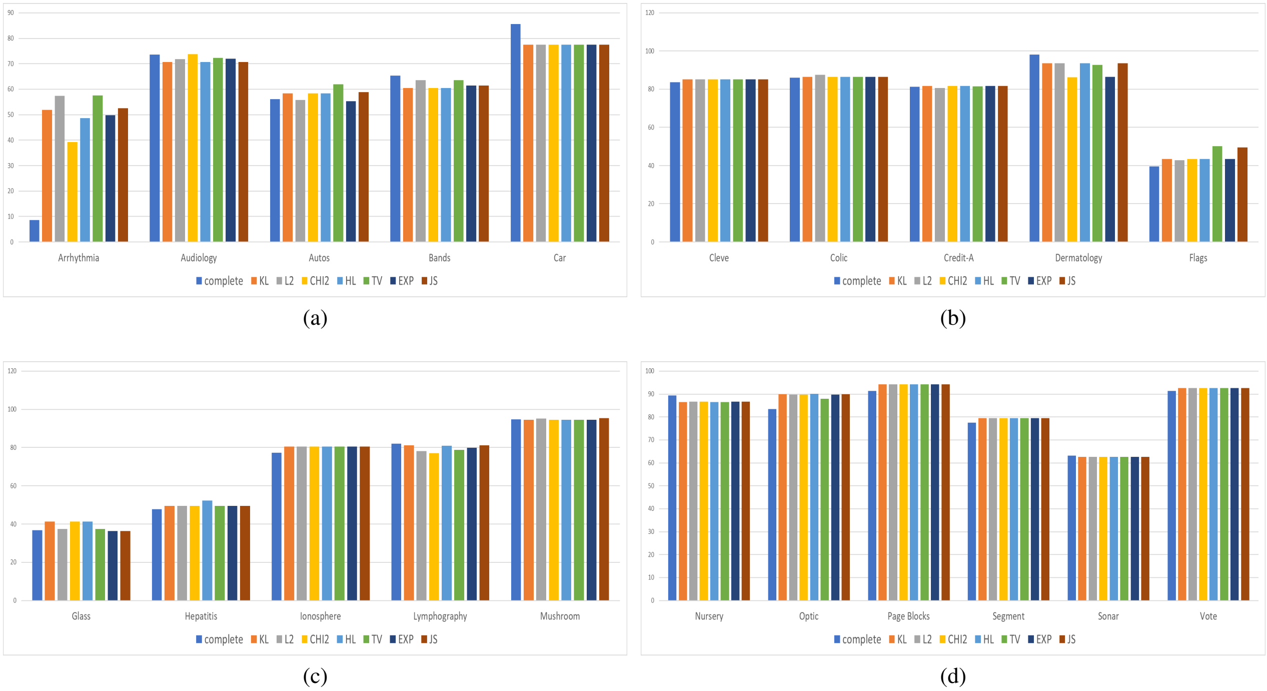
(c) (d)

Figure 11. F1-Score after each FS method with 75% threshold on all datasets

4.3. *F1-Score*

Figure 11 shows the F1-score results after each FS method with a 75% threshold is performed. Among all methods, EXP was the method that led to better results than all other methods. Win-lose-tie record of EXP against KL, L2, CHI2, HL, TV, and JS was 3-1-17, 7-3-11, 3-1-17, 6-2-13, 7-3-11, and 5-2-14, respectively. After EXP, CHI2 and KL led to better F1-score percentages than other remaining methods. These methods had a winning record against all the other methods except for EXP. After these three methods, HL, JS, TV, and L2 were ranked respectively. Also, EXP, CHI2, and KL in 13 out of 21 datasets led to higher F1-score percentages than the case of a total set of features. The worst records in this comparison jointly belong to L2 and JS with only 11 wins against the total set of features. Furthermore, EXP, KL, CHI2 led to the highest increase in F1-score on average among all methods, with approximately 1.27%, 1.25%, and 1.21% respectively. The lowest increase of F1-score on average was led by JS with 0.44%.

Results for F1-score percentage with a 50% threshold as a cutoff point were somewhat similar to the accuracy percentage results in Figure 10. Figure 12 shows the F1-score percentage results for all methods on 21 different datasets. As can be seen, JS is concluded to give better results than all other methods, with win-lose-tie records of 8-2-11, 8-3-10, 9-4-8, 6-6-9, and 7-1-13 against KL, L2, CHI2, HL, TV, and EXP. After JS, KL gained better results compared with other remaining methods. Its win-lose-tie records against L2, CHI2, HL, TV, EXP and JS was 6-6-9, 4-2-15, 3-2-16, 6-5-10, 7-3-11, and 2-8-11 respectively. After KL, HL, L2, TV, CHI2, and EXP were ranked respectively. In comparison with the total set of features, CHI2 recorded 14-7-0 which was better than all other methods, although CHI2 did not rank high in comparison with other FS methods. Also, KL, HL, TV, and JS recorded 13-8-0, L2 recorded 12-9-0 and at last, EXP concluded 11-10-0. Furthermore, the highest increase of F1-score percentage on average was led by TV with 2.87% and after that, JS with 2.54%. The lowest increase in F1-score percentage on average was led by CHI2 with 1.33%.



(a) (b)

(c) (d)

Figure 12. F1-Score after each FS method with 50% threshold on all datasets

5. Conclusion

Feature selection is a way of dimension reduction that selects a suitable subset of features. Filter feature selection methods, do not depend on which classification technique will be used and their procedure only includes data and its qualities. Divergence functions in statistics are measures of divergence between two or more probability distribution functions, and in a way, can be used as feature selection methods. In this paper, direct usage of several ranker divergence-based feature selection methods was studied and evaluated. Two thresholds were chosen to select the suitable subset of features from a ranked set of features based on each method. After performing classification, the results were observed and compared in order to find which FS methods perform better under which circumstances. EXP, CHI2, and JS were shown to perform better than the others while maintaining good results and leading to little to zero accuracy loss compared with the complete set of features.

The main limitation of divergence-based feature selection methods is their inability to explore redundancy amongst features, as they only take the relevancy of features into account. An optional future work would be dealing with this limitation and also evaluation of performances of these methods with considering different kinds of classifiers.

REFERENCES

1. S.I. Amari, *α-Divergence Is Unique, Belonging to Both f-Divergence and Bregman Divergence Classes*, IEEE Transactions on Information Theory, 55(11), pp.4925–4931, 2009.
2. M. Basseville, *Divergence measures for statistical data processingAn annotated bibliography*, Signal Processing, 93(4), pp.621–633, 2013.
3. L.M. Bregman, *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming*, USSR: computational mathematics and mathematical physics, 7(3), pp.200–217, 1967.
4. O.Calin, and C. Udriste, *Geometric modeling in probability and statistics*, Berlin: Springer, 2014.
5. A. Cichocki, and S.I. Amari, *Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities*, Entropy, 12(6), pp.1532–1568, 2010.
6. L. Cui, Y. Jiao, L. Bai, L. Rossi, and E.R. Hancock, *Adaptive feature selection based on the most informative graph-based features*, In International Workshop on Graph-Based Representations in Pattern Recognition, pp. 276–287. Springer, Cham, 2017.
7. D. Dua, and C. Graff, *UCI Machine Learning Repository . Irvine, CA: University of California, School of Information and Computer Science*, [http://archive.ics.uci.edu/ml], 2019.
8. G.H. Fu, Y.J. Wu, M.J. Zong, and J. Pan, *Hellinger distance-based stable sparse feature selection for high dimensional class-imbalanced data*, BMC bioinformatics, 21, pp.1–14, 2020.
9. B. Fuglede, and F. Topsoe, *Jensen-Shannon divergence and Hilbert space embedding*, In International Symposium on Information Theory, ISIT Proceedings. p. 31, IEEE, 2004, June.
10. I. Guyon, A. Elisseeff, *An introduction to variable and feature selection*, Journal of machine learning research, 3 (Mar), pp.1157–1182, 2003.
11. I. Guyon, A. Elisseeff, *An introduction to feature extraction*, In Feature extraction, pp.1–25. Springer, Berlin, Heidelberg, 2006.
12. R. Guzmn-Martnez, and R. Alaiz-Rodrguez, *Feature selection stability assessment based on the jensen-shannon divergence*, In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 597–612. Springer, Berlin, Heidelberg, 2011, September.
13. E. Hart, K. Sim, B. Gardiner, and K. Kamimura, *A hybrid method for feature construction and selection to improve wind-damage prediction in the forestry sector*, In Proceedings of the Genetic and Evolutionary Computation Conference, pp.1121–1128, 2017, July.
14. A. Hashemi and M.B. Dowlatshahi, *MLCR: a fast multi-label feature selection method based on K-means and L2-norm*, In 2020 25th International Computer Conference, Computer Society of Iran (CSICC), pp.1–7, IEEE, 2020, January.
15. Y. Jiang, N. Zhao, L. Peng, and S. Liu, *A new hybrid framework for probabilistic wind speed prediction using deep feature selection and multi-error modification*, Energy Conversion and Management, 199, p.111981, 2019.
16. S. Kullback, and R.A. Leibler, *On information and sufficiency*, The annals of mathematical statistics, 22(1), pp.79–86, 1951.
17. V. Kumar, and S. Minz, *Feature selection: a literature review*, SmartCR, 4(3), pp.211–229, 2014.
18. T.N. Lal, O. Chapelle, J. Weston, and A. Elisseeff, *Embedded methods*, In Feature extraction , pp.137–165. Springer, Berlin, Heidelberg, 2006.
19. C. Lee, and G.G. Lee, *Information gain and divergence-based feature selection for machine learning-based text categorization*, Information processing & management, 42(1), pp.155–165, 2006.
20. J. Li, K. Cheng, S. Wang, F. Mostatter, R.P. Trevino, J. Tang, and H. Liu, *Feature selection: A data perspective*, ACM Computing Surveys (CSUR), 50(6), pp.1–45, 2017.
21. Y. Lifang, Q. Sijun, and Z. Huan, *Feature selection algorithm for hierarchical text classification using Kullback-Leibler divergence*, In IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCBDA), pp. 421–424, IEEE, 2017.
22. J. Lin, *Divergence measures based on the Shannon entropy*, IEEE Transactions on Information theory, 37(1), pp.145–151, 1991.
23. H. Liu, and R. Setiono, *Chi2: Feature selection and discretization of numeric attributes*, In Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence, pp. 388–391. IEEE, 1995, November.

24. K.R. Niazi, C.M. Arora, and S.L. Surana, *Power system security evaluation using ANN: feature selection using divergence*, Electric Power Systems Research, 69(2-3), pp.161–167, 2004.
25. J. Novovicov, P. Pudil, and J. Kittle, *Divergence based feature selection for multimodal class densities*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 18(2), pp.218–223, 1996.
26. F.S. sterreicher, *Csiszrs f-divergences-basic properties*, RGMIA Res. Rep. Coll, 2002.
27. J.R. Quinlan, *Induction of decision trees*, Machine learning, 1(1), pp.81–106, 1986.
28. K.M. Schneider, *A new feature selection score for multinomial naïve Bayes text classification based on KL-divergence*, In Proceedings of the ACL Interactive Poster and Demonstration Sessions, pp. 186–189, 2004, July.
29. F. Thabtah, F. Kamalov, S. Hammoud, and S.R. Shahamiri, *Least Loss: A Simplified Filter Method for Feature Selection*, Information Sciences, 2020.
30. P. Temrat, Y. Jiraraksopakun, Y. Bhatranand, and K. Wea-asae, *Suitable feature selection for OSA classification based on snoring sounds*, In 2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), pp.1–4, IEEE, 2018.
31. T. Van Erven, and P. Harremos, *Rnyi divergence and Kullback-Leibler divergence*, IEEE Transactions on Information Theory, 60(7), pp.3797–3820, 2014.
32. J. Wang, Z. Feng, N. Lu, and J. Luo, *Toward optimal feature and time segment selection by divergence method for EEG signals, classification*, Computers in biology and medicine, 97, pp.161–170, 2018.
33. S. Yoon, Y. Song, K.C. Bureau, M. Kim, F.C. Park, and Y.k. Noh, *Interpretable Feature Selection Using Local Information For Credit Assessment*, 2018.
34. Y. Zhang, S. Li, T. Wang, and Z. Zhang, *Divergence-based feature selection for separate classes*, Neurocomputing, 101, pp.32–42, 2013.