A Comparative Analysis of Feature Selection Stability Measures

Mohana Chelvan P,
Assistant Professor,
Department of Computer Science,
Hindustan College of Arts and Science,
Chennai – 603 103, India,
E-mail: pmohanselvan@rediff.com

Dr. Perumal K,
Associate Professor,
Department of Computer Applications,
Madurai Kamaraj University,
Madurai – 625 021, India,
E-mail: perumalmala@gmail.com

***Abstract*—Now-a-days, data mining become indispensable for business organizations for decision making, which makes use of the information from huge amount of archived data. Due to the advancements in information technology, there will be proliferation of extremely high-dimensional data. Feature selection manages the "curse of dimensionality" as it is an important dimensionality reduction technique. Recently, Stability or robustness of feature selection methods becomes a hot topic of interest for researchers. Feature selection stability is the measure of the sensitivity of feature selection algorithms for the slight perturbations in the experimental dataset. There are various selection stability measures which have been used to measure the stability of feature selection algorithms based on the result sets. This paper gives an account of various selection stability measures and also the merits and demerits of each stability measure that have been explored using a set of experimental datasets.**

Keywords*—*selection stability; feature selection; stability measure; data mining

I. INTRODUCTION

The harvested data from organizational day-to-day activities become high dimensional with huge sample size due to the growth of high-throughput technologies. Data mining is very important for getting hidden knowledge from the huge datasets. The data mining techniques are mostly useless due to the high level of noise associated with collected samples on the non pre-processed data. Dimensionality reduction is the technique to remove the noisy and redundant attributes. Feature selection will selects small subset of relevant features and is an important dimensionality reduction technique. The benefits of feature selection methods will be reducing computation time, improving classification performance and a better understanding of the data in data mining applications [1].

The feature selection stability is the robustness of feature selection algorithm by which the subsequent iterations of feature selection must select same or similar subset of features and otherwise it will create confusion in the researchers mind. There are various measures to quantify the selection stability which are called as selection stability measures. The selection stability measurement should have [2] monotonicity, limits and correction for chance. Monotonicity is due to the large overlap between selected subsets and so the result should be in big stability values. Limits should bind each stability assessment method's result between constants such as [0, 1] or [–1, 1]. High dimensionality of the selected subset will create higher intersection by chance. Each measurement should have correction for chance as a constant that should correct the result in case of intersection by chance. The feature selection stability is mostly depends on the characteristics of the dataset [3] [4].

II. CATEGORIES OF FEATURE SELECTION STABILITY MEASURES

Stability can be assessed by the pairwise comparison between the resulting subsets obtained by feature selection algorithm on datasets. The stability is higher if the similarity between the resulting subsets is greater. Based on the output of the feature selection method, the stability measures are of three different representations i.e., indexing, ranking, and weighting [5].

A. Stability by Index

In this category of measurements, the selected subset of features is represented as a binary vector with cardinality equal to the total number of features m or as a vector of indices relating to the selected features k. Unlike the other stability measurements i.e., rank or weight based measurements, the index measurements have the possibility for handling subset of features, i.e., the number of selected features $k \leq m$. The index measurements assess the amount of overlap between the resulting subsets of features for assessing the stability. The examples for stability by index measurement are Dice's Coefficient, Tanimoto Distance, Jaccard Index and Kuncheva Index.

B. Stability by Rank

The stability by rank method assesses the stability by evaluating the correlation between the ranking vectors. Unlike the index method, these methods do not deal with partial set of features as they cannot handle vectors with different cardinality i.e., vectors

that correspond to different set of features. The measurements in this category include Spearman's Rank Correlation Coefficient SRCC.

C. Stability by Weight

Similar to the stability by rank, this category of measurement deals with only full subset of features. This method assesses selection stability by evaluating the weight of the full feature set. The stability by weight category of measurement has only one member called the Pearson's Correlation Coefficient PCC. Here the stability is assessed by evaluating the correlation between the two sets of weights w_i and w_j for the entire feature set in the dataset.

III. FEATURE SELECTION STABILITY MEASURES

A. Dice's Coefficient

Dice, Tanimoto and Jaccard are similar stability measures. Dice coefficient calculates selection stability by evaluating the overlap between two subsets of features as in (1) and is used in [6].

$$\text{Dice}(\mathcal{F}'_1, \mathcal{F}'_2) = \frac{2\,|\mathcal{F}'_1 \cap \mathcal{F}'_2|}{|\mathcal{F}'_1| + |\mathcal{F}'_2|} \quad (1)$$

Dice bounds between the values of 0 and 1, where 0 means the results are unstable i.e., no overlap between the subset of features and 1 means the two subsets are stable or identical.

B. Tanimoto Distance

Tanimoto measures selection stability by evaluating the amount of overlap between two subsets of the dataset as in (2) and produces value that bonds between the range of 0 and 1 [7].

$$\text{Tanimoto}(\mathcal{F}'_1, \mathcal{F}'_2) = 1 - \frac{|\mathcal{F}'_1| + |\mathcal{F}'_2| - 2\,|\mathcal{F}'_1 \cap \mathcal{F}'_2|}{|\mathcal{F}'_1| + |\mathcal{F}'_2| - |\mathcal{F}'_1 \cap \mathcal{F}'_2|} \quad (2)$$

C. Jaccard Index JI

The given different results $R = \{R_1, R_2 \ldots R_l\}$ will correspond to l different folds of the sample dataset. By evaluating the amount of overlap between the subsets in R, the stability can be assessed as in (4). By evaluating the amount of overlap between the features of the selected subsets, the JI is to assess the stability for subsets of features that contain indices of selected features [7]. The similarity between finite numbers of subsets is measured by the Jaccard coefficient measures. JI is measured as the size of the intersection of the selected subsets of features divided by the size of their union. JI for two selected subsets is shown by (3) and for a number of subsets in subsequent iterations is shown by (4).

$$S_J(R_i, R_j) = \frac{|R_i \cap R_j|}{|R_i \cup R_j|} \quad (3)$$

$$S_J(R) = \frac{2}{l\,(l\text{-}1)} \sum_{i=1}^{l\text{-}1} \sum_{j=i+1}^{l} S_J(R_i, R_j) \quad (4)$$

The Jaccard Index S_J returns a value which bounds in the interval of [0, 1] where 0 means the two subsets R_i and R_j of feature selection results are not stable and overlapped and 1 means the results are very stable and identical.

D. Kuncheva Index KI

In most of the stability measures, there will be overlap between the two subsets of the features due to chance. The larger cardinality of the selected features' lists positively correlated with the chance of overlap. To overcome this drawback, the Kuncheva Index KI which is proposed in [2] contains correction term to avoid the intersection by chance as in (5). KI is the only measurement that obeys all the requirements appeared in [2] i.e., Monotonicity, Limits and Correction for chance. The correction for chance term was introduced in KI and so it becomes desirable. So, unlike the other measurements, the larger value of cardinality will not affect the stability value.

$$\text{KI}(\mathcal{F}'_1, \mathcal{F}'_2) = \frac{|\mathcal{F}'_1 \cap \mathcal{F}'_2| \,.\, m - k^2}{k\,(m - k)} \quad (5)$$

KI's results bounds between the ranges of [–1, 1], where –1 means k = m/2, i.e., there is no intersection between the two subsets of features. KI becomes 1 when the cardinality of the intersection set equals k, i.e., $\mathcal{F}'_1$ and $\mathcal{F}'_2$ are identical. KI becomes close to zero for dissimilarly drawn lists of subset of features.

E. Spearman's Rank Correlation Coefficient SRCC

It is stability by rank method and it assesses the stability by evaluating two ranked subsets of features' r and r′, A. Kalousis et al. have used Spearman's Rank Correlation Coefficient in [5] and is given in (6).

$$\text{SRCC}(r, r') = 1 - 6 \sum_t \frac{(r_t - r'_t)^2}{m\,(m^2 - 1)} \quad (6)$$

The Spearman's results will bounds between the range of [–1, 1]. The result becomes 1 when the two ranks of the features' lists r and r′ are identical while it becomes –1 when ranks are exactly in inverse order and 0 means no correlation at all.

F. Pearson's Correlation Coefficient PCC

The PCC is the only stability measure in the stability by weight category and is used in [5]. It is a symmetric measure and it measures the correlation between the weights of the subset of features w and w′ as in (7).

$$\text{PCC}(w, w') = \frac{\sum_i (w_i - \mu_w)(w'_i - \mu_{w'})}{\sqrt{\sum_i (w_i - \mu_w)^2 \sum_i (w'_i - \mu_{w'})^2}} \quad (7)$$

Here μ is the mean of the features weight. PCC takes values between –1 and 1, where –1 means that the subsets features weights w and w′ are anti-correlated and 1 mean the weight vectors of features' lists are perfectly correlated, while 0 means no correlation. When the weight is equal to zero for large number of features, the stability will be shown higher. Even though, the algorithm assigns weight within the bounds of 1 and –1 and so this will not be an issue.

IV. FEATURE SELECTION ALGORITHMS

The feature selection process is mostly based on three approaches viz. filter, wrapper and hybrid [8]. The filter approach of feature selection is by removing features on some measures or criteria and the goodness of a feature is evaluated using intrinsic or statistical properties of the dataset. A feature is adjudged as the most suitable feature based on these properties, and is selected for machine learning or data mining applications [9]. In the wrapper approach, the subset of features is generated and then goodness of the subset is evaluated using some classifier. The purpose of some classifier in this approach is to rank the features in the dataset and based on this rank, a feature is selected for the required application. The embedded model combines the advantages of both the above models. The hybrid approach takes advantage of the two approaches by exploiting the different evaluation criteria of them in different search stages.

A. Correlation-based Feature Selection CFS

The worth of a subset of attributes is evaluated by CFS by considering the degree of redundancy between them along with the individual predictive ability of each feature. Subsets of features that are having low inter-correlation between the classes but that are highly correlated within the class are preferred [10]. CFS determines the best feature subset and can be combined with search strategies such as genetic search, best-first search, backward elimination, forward selection and bi-directional search. Authors have GA as search method with CFS as fitness function.

$$r_{zc} = \frac{k\, r_{zi}}{\sqrt{k + (k-1)\, r_{ii}}} \qquad (8)$$

CFS is given in (8) where r_{zc} is the correlation between the class variable and the summed subset features, k is the number of subset features, r_{zi} is the average of the correlations between the class variable and the subset features and r_{ii} is the average inter-correlation between subset features [10].

B. Information Gain IG

The entropy is a criterion of impurity in a training set *S*. It is defined as a measure reflecting additional information about Y provided by X that represents the amount by which the entropy of Y decreases [11]. This measure is known as IG as in (9).

$$IG = H(Y) - H(Y/X) = H(X) - H(X/Y) \qquad (9)$$

IG is a symmetrical measure as the information gained about X after observing Y is equal to the information gained about Y after observing X. IG has the weakness that it is biased in favour of features with more values even when they are not more informative. The information gain with respect to the class is measured by which the worth of an attribute is evaluated. The independence between a feature and the class label is assessed by IG by considering the difference between the entropy of the feature and the conditional entropy given by the class label as in (10).

$$IG\,(Class, Attribute) = H\,(Class) - H\,(Class \mid Attribute) \qquad (10)$$

V. EXPERIMENTAL RESULTS

The five datasets used in the experiments are splice, connect-4, kddcup, optdigits and spambase. The datasets are obtained from the KEEL dataset repository [12]. Table 1 shows the characteristics of the datasets. In the listed datasets, the splice and connect-4 have categorical values. The kddcup dataset has both categorical and numeric values. The optdigits dataset has only integer values and spambase dataset has only real values.

CFS (Correlation-based Feature Selection) feature selection algorithm was used for Dice, Tanimoto, Jaccard and Kuncheva measures while Information Gain feature selection algorithm was used for Spearman and Pearson measures. Greedystepwise

TABLE 1. CHARACTERISTICS OF DATASETS SPLICE, CONNECT-4, KDDCUP, OPTDIGITS AND SPAMBASE

S. No.	*Datasets Characteristics*	*Datasets*				
		Splice	***connect-4***	***kddcup***	***optdigits***	***spambase***
1	Type	Classification	classification	Classification	classification	classification
2	Origin	Real World	Real World	Real World	Real World	Real World
3	Instances	3190	67557	494020	5620	4597
4	Features	60	42	41	64	57
5	Classes	3	3	23	10	2
6	Missing Values	No	No	No	No	No
7	Attribute Type	Nominal	Nominal	Real / Nominal	Integer	Real

TABLE 2. SELECTION STABILITY MEASURE VALUES FOR THE DATASETS SPLICE, CONNECT-4, KDDCUP, OPTDIGITS AND SPAMBASE

Stability Measures	*Bounds*	*Datasets*				
		Splice	*Connect-4*	*kddcup*	*optdigits*	*spambase*
Dice	[0,1]	0.8372093	0.66666667	0.72727273	0.95890411	0.77419355
Tanimoto	[0,1]	0.72	0.5	0.57142857	0.92105263	0.63157895
Jaccard	[0,1]	0.72	0.5	0.57142857	0.92105263	0.63157895
Kuncheva	[-1,1]	0.72272727	0.51827243	0.62727273	0.9375	0.72380952
Spearman's	[-1,1]	1	1	1	1	1
Pearson's	[-1,1]	0.967828	0.838915	0.986779	0.995748	0.909666

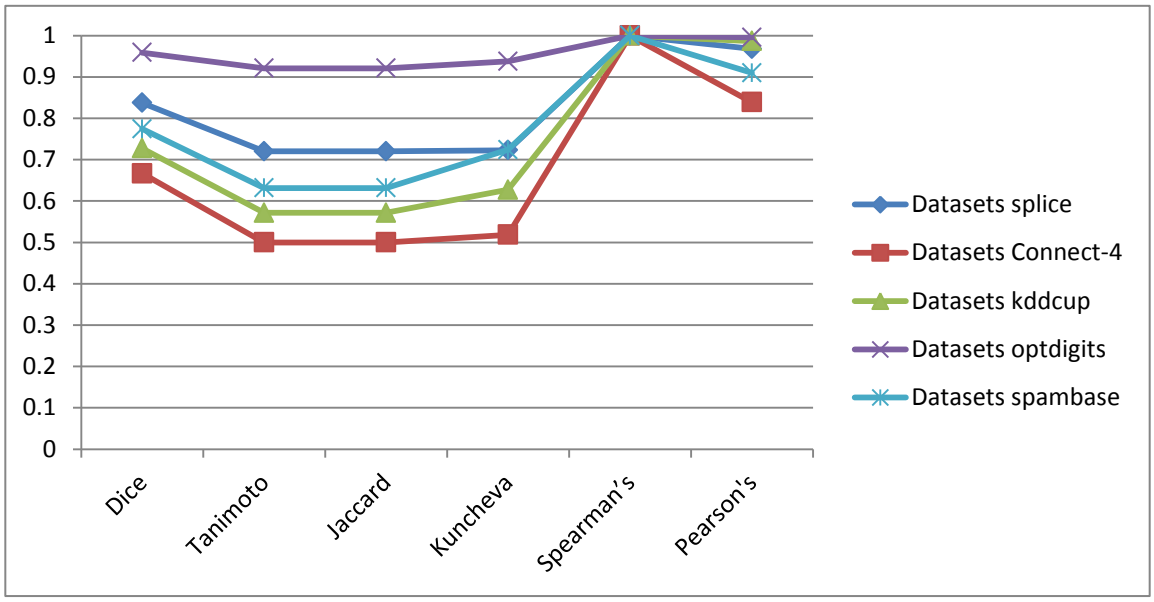

Fig. 1. Graph showing the selection stability measure values for the five datasets splice, connect-4, kddcup, optdigits and spambase

was the attribute evaluator for CFS and Ranker was the attribute evaluator for Information Gain.

GreedyStepwise performs a greedy backward or forward search through the space of attribute subsets. It may start with all/no attributes or from an arbitrary point in the space. It stops when the deletion/addition of any remaining attributes results in a decrease in evaluation. By recording the order that the attributes are selected in traversing the space from one side to the other, the ranked list of attributes can also produced by it. Attributes are ranked by their individual evaluations in the case of ranker attribute evaluator.

Table 2 gives the stability measure values for the five sample datasets. The Dice, Tanimoto, and Jaccard measures have similarity in their results. Dice gives slightly higher stability results which equals to the exact amount of overlap with respect to the intersection between the two subsets. However, in the case of Tanimoto and Jaccard index measures, they become closer to m because overlaps by chance are higher values when the subsets cardinalities get higher. In case of intersection by chance, they don't have constant to correct in contrasting with the measure Kuncheva Index. When k gets larger and closer to m, they will gives higher stability values due to the impact of the number of selected features k on the stability. However, an advantage of these measurements in comparing with other categories i.e., stability by rank and stability by weight, they can deal with different number of subsets of features, i.e., of different cardinalities. They do not take the dimensionality m in the measurement but they take into account the number of selected features k. But in the case of Kuncheva Index, correction term gives negative weight to k and so it does not suffer from the same drawback.

From the Fig. 1, it has been shown that Tanimoto Distance and Jaccard Index have the same values for all datasets and the Spearman's Rank Correlation

Coefficient SRCC will have the maximum values for all the datasets.

VI. CONCLUSION

The feature selection stability is mostly depends on the characteristics of dataset but is also not completely independent on the feature selection algorithms. Selecting suitable selection stability measure for the given dataset and feature selection algorithm is an interesting research problem. This paper gives an account of important feature selection stability measures and their characteristics. From the listed stability measures, Kuncheva Index will give reasonably meaningful values because of the correction of chance term.

REFERENCES

[1] A. K. Jain & B. Chandrasekaran, Dimensionality and sample size considerations in pattern recognition practice. In P. R. Krishnaiah & L. N. Kanal (Eds.), Handbook of Statistics, 835–855, 1982

[2] L. I. Kuncheva, A stability index for feature selection, In Proceedings of the 25th conference on Proceedings of the 25th IASTED International Multi Conference: artificial intelligence and applications, pages 390 - 395, Anaheim, CA, USA, 2007, ACTA Press

[3] Salem Alelyani, Huan Liu, The Effect of the Characteristics of the Dataset on the Selection Stability, 1082-3409/11 IEEE DOI 10.1109/International Conference on Tools with Artificial Intelligence.2011.167, 2011

[4] Salem Alelyani, Zheng Zhao, Huan Liu, A Dilemma in Assessing Stability of Feature Selection Algorithms, 978-0-7695-4538-7/11, IEEE DOI 10.1109/ International Conference on High Performance Computing and Communications. 2011.99, 2011

[5] A. Kalousis, J. Prados, and M. Hilario, Stability of feature selection algorithms: a study on high-dimensional spaces, Knowledge and Information Systems, 12(1):95 - 116, May 2007

[6] L. Yu, C. Ding, and S. Loscalzo, Stable feature selection via dense feature groups, In KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 803 - 811, New York, NY, USA, 2008. ACM.

[7] Y. Saeys, T. Abeel, and Y. Van de Peer, Robust feature selection using ensemble feature selection techniques, 2008.

[8] K. Sudha, J. JebamalarTamilselvi, "A Review of Feature Selection Algorithms for Data Mining Techniques", International Journal on Computer Science and Engineering (IJCSE) ISSN: 0975-3397, Vol. 7, No.6, pp. 63-67, June 2015

[9] K. Mani, P. Kalpana, "A review on filter based feature selection", International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE) ISSN: 2320-9801, Vol. 4, Issue 5, May 2016

[10] Mark A Hall, "Correlation-based Feature Selection for Machine Learning", Dept of Computer science, University of Waikato, 1998, http://www.cs.waikato.ac.nz/ mhall / thesis.pdf

[11] Hall M A and Smith L A, "Practical feature subset selection for machine learning", Proceedings of the 21st Australian Computer Science Conference, 181– 191, 1998

[12] Alcalá-Fdez J, Fernández A, Luengo J, Derrac J, García S, Sánchez L and Herrera F, "KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," J. Multiple-Valued Logic Soft Comput., 17(2): 255–287, 2010