A Survey on Feature Selection Stability Measures

Mohana chelvan P.
Dept. of Computer Science,
Hindustan College of Arts and Science,
Chennai – 603 103, India,
Email: pmohanselvan [AT] rediff.com

Perumal K.
Dept. of Computer Applications,
Madurai Kamaraj University,
Madurai – 625 021, India

Abstract - **Data mining is the extraction of useful knowledge from stored data of organizations which can be used for decision making. Feature selection is an important dimensionality reduction technique for high dimensional datasets which is used in data mining. In recent years, feature selection stability becomes the hot topic of research in data mining. Stability becomes an important criterion for feature selection algorithms. While studying selection stability, there is also need to analyse stability measures. There are numerous ways to measure selection stability. This paper gives an overview of different categories of stability measures and also gives an account of important stability measures in each category.**

Keywords - data mining, feature selection, selection stability, stability measure

I. INTRODUCTION

Data mining is indispensable for business organizations for getting edge over their competitors. Feature selection is important technique of data mining for dimensionality reduction. Selection stability becomes important topic of research for feature selection algorithms. The stability of the feature selection methods is defined as the sensitivity of the feature selection algorithms to the small variation or perturbation of the dataset. The stability measures have been categorised in three main categories based on the representation of the output of the selection method [7]. These categories are stability by index, stability by rank and stability by weight. In stability by index, the indices of the selected features are considered. In this category, the selected features will not have particular order or corresponding relevant weight. In the category of stability by rank, the ranked list of the features will influence in stability evaluation. In stability by weight category, each feature is assigned a weight according to the degree of relevance.

All the three categories of stability are generated from the features' weights. However, each domain will be interested in different output and will be interested in particular category of stability of that particular output. It is important to emphasize that same rank does not necessarily having the same weight and same selected subset. The three requirements that each stability measurement should have [8] are as follows:

- Monotonicity: If there is large overlap between selected subsets, the result should be in large stability.
- Limits: Each stability assessment method's result should be bounded between constants such as [-1, 1] or [0, 1]. These bounds are independent of any dataset factor including the dimensionality of the dataset m or the number of selected features k. These limits should be at minimum when the sets are completely unstable and become at maximum when they become identical or stable.
- Correction for chance: The measurement should have a constant that correct the result in case of intersection by chance occur due to high dimensional selected subset. The larger the cardinality of selected subsets then there is more chance for larger intersection between subsets.

Most of the measures will not consider all the above three requirements but the only measurement that considers all the three requirements is Kuncheva Index (KI) [8]. In addition to these requirements, there are some important properties that should be taken into consideration due to their impact on the stability result [1] [2]. These properties include:

- The dimensionality of the dataset m. It is an important factor that may affect the stability of an algorithm.
- The number of selected features k. These two factors implicitly mentioned in the correction for chance requirement. However, they should be considered in other ways too. For example, these two factors, i.e. m and k, are considered in order to rank two algorithms in terms of the stability.
- The sample size n. It has significant impact on the stability. These three factors will be considered in justifying the differences in the stability of algorithms.
- The data variance. It was demonstrated in [5] that the data variance has a huge impact on the stability. To judge or compare algorithms in terms of stability, the variance of the dataset and perhaps other important underlying characteristics of the dataset are taken into consideration.

- The symmetry of the measurement. The stability value should not be sensitive to the order of the results.

Stability can be assessed simply by pairwise comparison between the results. Hence, the stability is higher if the similarity is greater. There are three different representations of the output of the feature selection methods, i.e., indexing, ranking, and weighting [7] and hence different measures should be fit in these representations. The remaining part of the paper will explain these measurements and others categorized by the output scheme.

II. STABILITY BY INDEX

In the stability by index category, the selected subset of features is represented as either a vector of indices corresponding to the selected features or as a binary vectors with cardinality equals m, where $f_i = 1$ means that the i^{th} feature is selected. In this category of measurements, there is the possibility of handling a number of selected features $k \leq m$ but it is not possible with the rank or weight measurements. Hence, these measurements have common result's limits where some in the interval [0, 1] and others in [-1, 1] while others are not bounded at all. This category of measurements assesses the amount of overlap between results in order to assess the stability. The important measurements in the category of stability by index are as follows:

A. Average Normal Hamming Distance (ANHD)

Average Normal Hamming Distance measure was used in [3] to assess the stability of feature selection algorithm which is for subset of selected features. ANHD measures the amount of overlap between two subsets. ANHD ($\hat{H}$) works with binary representation that represent the selected feature subset $\hat{f}_{ik}$, 1 and 0 indicate whether the k^{th} feature was selected in the i^{th} run or not, respectively.

$$\hat{H}(\hat{f}_i,\hat{f}_j) = \frac{1}{m}\sum_{k=1}^{m} \left| \hat{f}_{ik}-\hat{f}_{jk} \right|$$

When m becomes larger, $\hat{H}$ becomes smaller which will lead to more stable algorithm. When small numbers of features were selected, i.e. have values equal to 1, and the rest are set to zero, then $\hat{H}$ will be small. This is due to the fact that selected features across ℓ-folds will be treated as unselected ones. If a feature f_i is selected in all ℓ or not selected, there will have the same impact on the stability result. This property of ANHD will lead in most cases to wrong conclusion about the stability especially when $k \ll m$ where the majority of the features are not selected. ANHD results will be in the interval [0, 1], where 0 is the most stable and 1 means not stable at all. ANHD cannot deal with different sizes of selected features' sets in relation with capability. As ANHD does not have the correction for chance constant, the result will be misleading.

B. Dice's Coefficient

Dice coefficient is a similarity measure used in [11] to calculate the overlap between two sets. It was related to Jaccard index.

$$\text{Dice}(F'_1,F'_2) = \frac{2\left|F'_1 \cap F'_2\right|}{\left|F'_1\right| + \left|F'_2\right|}$$

Dice takes value between 0 and 1, where 0 means no overlap and 1 means the two sets are identical. There will be similarity between the measures Dice, Tanimoto and Jaccard.

C. Tanimoto Distance and Jaccard's Index

Tanimoto is similar to Dice and it measures the amount of overlap between two datasets and produces value in the range 0 and 1.

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

It is easy to proof that Tanimoto is equivalent to Jaccard's index [9]:

$$\text{Jaccard}(F'_1, F'_2) = \frac{\left|F'_1 \cap F'_2\right|}{\left|F'_1 \cup F'_2\right|}$$

Although Dice, Tanimoto, and Jaccard behave similarly in all cases, Dice sometimes give slightly higher and more meaningful stability results with respect to the intersection between the two subsets. For example, assume that there will be two selected subsets with equal length, k = 20, and they intersect in 10 features, which is exactly 50% of total number of features for each set. Dice is going to give a stability equals to this exact amount of overlap i.e., 0.5, but Tanimoto and Jaccard are going to be 0.33 for each of them due the fact that they divide by the length of union of the two selected sets. These three measurements will give higher values when the subsets cardinalities get closer to m

because more overlaps by chance are higher. Hence, they don't have constant to correct in case of intersection by chance. An advantage of these measurements, unlike ANHD, they can deal with sets of different cardinalities. They do not take the dimensionality m in account but they comprise the number of selected features k in the measurement.

D. Kuncheva Index KI

The drawback of most stability measurements is that the larger the cardinality of the selected features' lists, the more overlap between lists due to chance. To overcome this drawback, [8] proposed Kuncheva Index KI that contains correction term to avoid intersection by chance between the two subsets of the features.

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

KI's results ranges [-1, 1], where -1 means there is no intersection between the lists and k = m/2. KI achieves 1 when $\mathcal{F}'_1$ and $\mathcal{F}'_2$ are identical which means the cardinality of the intersection set equals k. KI values becomes close to zero for independently drown lists. KI is the only measurements that obey all the requirements appeared in [8]. KI becomes desirable because of the correction for chance term that was introduced in [8]. In the case of KI, larger value of cardinality will not affect the stability, but in other measurements, the larger the cardinality is, the higher the stability will be. In Jaccard Index, there will be impact of the number of selected features k on the stability. It gives higher stability values when k gets larger and closer to m. But KI does not suffer from the same drawback because the correction term gives negative weight to k.

E. Percentage of Overlapping Gene (POG)

POG is similar to Tanimoto and Jaccard measures and it is used to measure the consistency of the feature subsets by counting the amount of overlap between them. POG is not symmetric and hence $POG(\mathcal{F}'_1,\mathcal{F}'_2)$ is not necessarily equal to $POG(\mathcal{F}'_2,\mathcal{F}'_1)$, which is undesirable property in general. However, it will be symmetric if $|\mathcal{F}'_1| = |\mathcal{F}'_2|$. [12] proposed a matrix that introduced a new variable z into POG and that consider the correlated molecular changes in a biological data set. [12] defined POGR as the percentage of overlapping genes, or features, related matrix to evaluate the consistence between two differentially expressed genes lists.

$$POG(\mathcal{F}'_1,\mathcal{F}'_2) = \frac{|\mathcal{F}'_1 \cap \mathcal{F}'_2|}{|\mathcal{F}'_1|}$$

$$POGR(\mathcal{F}'_1,\mathcal{F}'_2) = \frac{|\mathcal{F}'_1 \cap \mathcal{F}'_2| + z}{|\mathcal{F}'_1|}$$

Where z represents the number of genes in $\mathcal{F}'_1$ that are not in $\mathcal{F}'_2$ but they are significantly positively correlated to at least one gene in $\mathcal{F}'_2$. By having z there will be overcoming one drawback of the previous measurements. All previous measurements ignore the redundancy or the correlation between the values of the features. In previous measures, including POG, these two features no way to be counted positively toward the stability. In other words, f_i and f_j won't be considered as one feature even if they are redundant or positively highly correlated. However, by introducing z, there will be able to capture the correlation between the features and, thus, consider such features as one single feature. [12] introduced a new matrix normalized version for POG and POGR, or nPOG and nPOGR for short, to overcome the dependency between the result and the list length by introducing the expected of the shared features $E(|\mathcal{F}'_1 \cap \mathcal{F}'_2|)$. In addition, they introduced the expected number of z, E(z) onto the POGR, as follow:

$$nPOG(\mathcal{F}'_1,\mathcal{F}'_2) = \frac{|\mathcal{F}'_1 \cap \mathcal{F}'_2| - E(|\mathcal{F}'_1 \cap \mathcal{F}'_2|)}{|\mathcal{F}'_1| - E(|\mathcal{F}'_1 \cap \mathcal{F}'_2|)}$$

$$nPOGR(\mathcal{F}'_1,\mathcal{F}'_2) = \frac{|\mathcal{F}'_1 \cap \mathcal{F}'_2| + z - E(|\mathcal{F}'_1 \cap \mathcal{F}'_2|) + E(z)}{|\mathcal{F}'_1| - E(|\mathcal{F}'_1 \cap \mathcal{F}'_2|) - E(z)}$$

Where $E(|\mathcal{F}'_1 \cap \mathcal{F}'_2|)$ can be simply estimated by the average of the scores for arbitrary number of pairs of random lists of length $|\mathcal{F}'_1|$ and $|\mathcal{F}'_2|$ respectively. Similarly, E(z) can be estimated as the average number of features in the list $\mathcal{F}'_1$ which are not shared but significantly positively correlated with features in the other list $\mathcal{F}'_2$. These two parameters are the correction for chance terms in these two measurements. Finally, the limit of the results in POG measures family varies. POG and POGR are bounded by 0 and 1 while nPOG and nPOGR are in the interval [-1, 1] which make the latter obey Kuncheva requirements.

F. Consistency Measures

All the previous measures' are based on to assess the overlap between the subsets by pairwise comparison of the subsets. Hence the complexity will be equal to or greater than $O((m.(\ell^2 - \ell))/2)$, where ℓ is the number of subsets of selected features. [10] proposed three consistency measures which will be superior in complexity time by overcoming such shortcomings. These measures calculate stability by

taking the frequency of each selected feature. Here, each subset is processed only once to count the frequency of each selected feature and hence the complexity becomes O(k.ℓ). In the following three consistency measures, S as an input where $S = f_1, f_2, ...f_\ell$. Here x is the union of all subsets in S and t is the total frequency in S.

a) Consistency Measure C

$$C(S) = \frac{1}{|x|} \sum_{i}^{|x|} \frac{r_i - 1}{\ell - 1}$$

b) Weighted Consistency Measure CW

$$CW(S) = \sum_{t}^{|x|} \frac{r_i}{t} \frac{r_i - 1}{\ell - 1}$$

c) Relative Weighted Consistency Measure CW_{rel}

$$CW_{rel}(S, m) = \frac{m(t-z+\sum_{i}^{m} r_i(r_i-1)) - t^2 + z^2}{m(h^2+l(t-h)-z) - t^2 + z^2}$$

where r_t is the rate of occurrence, i.e. frequency, of feature f_i and z and h are t mod m and t mod ℓ respectively. CW_{rel} will show the amount of randomness in the feature selection process and it neither evaluates the amount of overlap between the subsets nor the frequency of the features. If CW_{rel} gives small number and the others give higher numbers. It may indicate drawback in the process of selecting the features. There are no preferable features or the methods overfit, etc. [10]. Instead, the consistency measure C can be rewritten in a less complex way to show some hidden properties of this measure. Here $t = \sum_{t}^{|x|} rt$ and by subtracting $|x|$ from both sides the following equation can be obtained:

$$t - |x| = \sum_{i}^{|x|} r_i - 1$$

$$\sum_{i} |x| (k-1) = (k-1)|x|$$

Since W is a constant, we can rewrite C(S) as follow:

$$C(S) = \frac{t - |x|}{(k-1)|x|}$$

G. Symmetrical Uncertainty SU

L. Yu et al in [11] and G. Gulgenzen et al in [4] used an entropy based nonlinear correlation called Symmetrical Uncertainty SU. It considers the similarity of the feature values and not features indices and hence it is different from the previous measures. SU is the correlation between the values of the selected features across different selected subsets hence it satisfies the nice and desirable property when evaluating stability. For example, consider that f_i and f_j to be duplicated feature which are considered as two different features when evaluating the stability and they were selected in F'_1 and F'_2 respectively. But, SU will consider them as a single feature by the values of the features. SU is symmetric because the information gain IG(fi | fj) = IG(fj | fi). SU is not bounded by any constants and is the undesirable property. The stability results can be normalized by considering the number of selected features k. The SU in [11] used it to calculate the similarity between two sets of feature groups but [4] used it as similarity measure between two sets of individual features.

$$SU(f_i, f_j) = 2\left[\frac{IG(f_i | f_j)}{H(f_i) + H(f_j)}\right]$$

Where f_i and f_j are i^{th} and j^{th} selected features and IG and H are the information gain and the entropy, respectively, as follows:

$$IG(f_i|f_j) = H(f_i - H(f_i|f_j))$$

$$H(f_i) = \sum_{x \epsilon f_i} p(x) \cdot lg_2(p(x))$$

$$H(f_i|f_j) = \sum_{y \epsilon f_j} p(y) \sum_{x \epsilon f_i} p(x|y) \cdot lg_2(p(x|y))$$

The similarity between two sets will be the average of SU for all unique pairs of i and j. The SU is the most expensive measure and the complexity is due to the expensive computations of IG for each unique pairs of selected features. It depends on the number of selected features k, where the worst case is when k = m. There is need to normalize, discretize, and center the datasets before computing the SU. It also makes it even more expensive.

III. STABILITY BY RANK

In the stability by rank method, the evaluation is by the correlation between the ranking vectors. These methods also deal with full set of features. In other words, they cannot handle vector that correspond to different set of features or vectors with different cardinality.

A. Spearman's Rank Correlation Coefficient SRCC

To evaluate the stability of two ranked sets of features' r and r′, A. Kalousis et al. in [7] adapted Spearman's Rank Correlation Coefficient.

$$SRCC(r, r') = 1 - 6 \sum_{t} \frac{(r_t - r'_t)^2}{m\,(m^2-1)}$$

The result of Spearman's will be in the range of [-1, 1]. The maximum will be achieved when the two ranks are identical while the minimum is when they exactly in inverse order and 0 means no correlation at all between r and r′.

B. Canberra Distance CD

Canberra Distance is the absolute difference between two rank sets and the generalized form is given by:

$$CD(r, r') = \sum_{t}^{N} \frac{\left| r_i - r'_i \right|}{r_i + r'_i}$$

CD only has a lower bound. The result depends on the number of features. Hence the CD will be larger for the higher value of m. Therefore, CD can be normalized by dividing by m in order to obtain results between 0 and 1. A weighted version of CD was proposed in [6] called WCD:

$$WCD^{(k+1)}(r, r') = \sum_{t=1}^{N} \frac{\left| \min\{r_t, k+1\} - \min\{r'_t, k+1\} \right|}{\min\{r_t, k+1\} - \min\{r'_t, k+1\}}$$

In WCD, the most important features are located in the top-k positions of the ranked list. Hence, the variation in the upper position of the list should be more relevant than those in the lower part [6]. WCD can be normalized by the number of features in the same way as CD.

IV. STABILITY BY WEIGHT

In this category of measurements, there is the deal with the weight of the feature set w. This category has only one measurement called the Pearson's Correlation Coefficient PCC. It takes two sets of weights w_i and w_j for the entire feature set in the dataset and return the correlation between them as the stability. In contrast with the stability by index, this category cannot deal with different subsets size or subset of features.

A. Pearson's Correlation Coefficient PCC

[7] uses Pearson's to measure the correlation between the weights of the features that returned from more than one run. The Pearson's Correlation Coefficient PCC stability will be as following:

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

Here μ is the mean. PCC takes values between -1 and 1, where 1 means the weight vectors are perfectly correlated, -1 means they are anti-correlated while 0 means no correlation. The stability will be shown higher when the weight is equal to zero for big number of features. However, this will not be an issue in situations as the algorithm assigns weight between 1 and -1. The PCC is a symmetric measure and is the only stability measure that handles feature weights.

V. MEASUREMENTS CATEGORIES

In Table 1, the stability measurements have been categorised based on four criteria i.e. Bounds, Symmetrical, Different size and Complexity. Most of the stability assessment methods have been clearly shown in the Table 1. The existing measures deals with one output scheme only but not with two or more different output schemes. The measures by rank and by weight are not able to handle different subset sizes but this property is common among all measurements belonging to the category by index. The running time complexities of these methods are more or less similar.

VI. CONCLUSION

This paper gives an account of various stability measures in each category. It also gives information about the capabilities of the stability measures. From this comparative study we can get information about the strength and weakness of each measure and their suitability for the required experiments on feature selection.

REFERENCES

[1] 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

[2] 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

[3] K. Dunne, P. Cunningham, and F. Azuaje, Solutions to instability problems with sequential wrapper-based approaches to feature selection, Technical Report TCD-CD-2002-28, Department of Computer Science, Trinity College, Dublin, Ireland, 2002.

[4] G.Gulgezen, Z. Cataltepe, and L. Yu, Stable and accurate feature selection, In ECML/PKDD (1), pages 455 - 468, 2009.

[5] Y. Han and L. Yu, A Variance Reduction Framework for Stable Feature selection, In 2010 IEEE International Conference on Data Mining, pages 206{215.IEEE, 2010.

[6] G. Jurman, S. Merler, A. Barla, S. Paoli, A. Galea, and C. Furlanello, Algebraic stability indicators for ranked lists in molecular profiling Bioinformatics, 24(2): 258 - 264, Jan 2008.

[7] 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.

[8] 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.

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

[10] P. Somol and J. Novovivcova, Evaluating the stability of feature selectors that optimize feature subset cardinality, 2010.

[11] 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.

[12] M. Zhang, L. Zhang, J. Zou, C. Yao, H. Xiao, Q. Liu, J. Wang, D. Wang,C. Wang, and Z. Guo, Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes, Bioinformatics, 25(13):1662 - 1668, Jul 2009.

TABLE 1. THE CATEGORIES OF THE EXISTING STABILITY MEASUREMENTS.

Results	**Measures**	**Capability**				**Reference**
		Bounds	***Symmetrical***	***Different Size***	***Complexity***	
Index	ANHD	[1,0]	Yes	Yes	$O((m.(l^2-l))/2)$	[3]
	Dice	[0,1]	Yes	Yes	$O((m.(l^2-l))/2)$	[11]
	Jaccard	[0,1]	Yes	Yes	$O((m.(l^2-l))/2)$	[9]
	KI	[-1,1]	Yes	Yes	$O((m.(l^2-l))/2)$	[8]
	Tanimoto	[0,1]	Yes	Yes	$O((m.(l^2-l))/2)$	[9]
	Consistency	[0,1]	Yes	Yes	O((k.l)/2)	[10]
	CW	[0,1]	Yes	Yes	O((k.l)/2)	[10]
	CW_{rel}	[0,1]	Yes	Yes	O((k.l)/2)	[10]
	POG	[0,1]	No	Yes	$O((m.(l^2-l))/2)$	[12]
	nPOG	[-1,1]	No	Yes	$O((m.(l^2-l))/2)+O(c)$	[12]
	POGR	[0,1]	No	Yes	$O((m.(l^2-l))/2)+O(c)$	[12]
	nPOGR	[-1,1]	No	Yes	$O((m.(l^2-l))/2)+O(c)$	[12]
	SU	$[0,\infty]$	Yes	Yes	$O((m.(l^2-l))/2)$	[11]
Rank	Spearman's	[-1,1]	Yes	No	$O((m.(l^2-l))/2)$	[7]
	CD	$[0,\infty]$	Yes	No	$O((m.(l^2-l))/2)$	[6]
	WCD	$[0,\infty]$	Yes	No	$O((m.(l^2-l))/2)$	[6]
Weight	Pearson's	[-1,1]	Yes	No	$O((m.(l^2-l))/2)$	[7]