Stability Comparison of Feature Selection Algorithms using Normalized Rank Transformation

Taufik Djatna
Hiroshima University
1-7-1 Kagamiyama Higashi Hiroshima
739-8521, Japan
Email: taufikdjatna@hiroshima-u.ac.jp

Yasuhiko Morimoto
Hiroshima University
1-7-1 Kagamiyama Higashi Hiroshima
739-8521, Japan
Email: morimoto@mis.hiroshima-u.ac.jp

***Abstract*— There are needs for evaluating rank order-based similarity between different classifiers in feature selection. Feature selection maps from data set to give further understanding about the importance of ranking in decision making within feature selection algorithms. The results are ordered rankings of training and testing data. In order to compare stability within each classifier, we deploy normalized rank transformation approach to get the degree of similarity between training and testing data set. The accuracy of the selected features is then evaluated using various classifiers.**

I. Introduction

Feature selection is one of the important and frequently used techniques in data preprocessing for data mining [1]. It reduces the number of features, removes irrelevant, redundant, or noisy data, and brings the immediate effects for applications. As results, we can speed up a data mining algorithms, improve mining performance such as predictive accuracy. Feature selection is a process that selects a subset of original features. The task of feature selection is to determine which features to select and deploy in order to predict a target attribute. Features are relevant if their value vary systematically with category membership. Feature selection algorithms have critical roles in many applications of machine learning, CRM, data mining and genomic analysis [2].

With these wide area of applications, there are needs for evaluating rank order-based on similarity between different classifiers in feature selection algorithms. The optimal subset can be viewed as the ranked order of the selected features. This work tries to evaluate the ranking stability among features selection algorithms by quantifying the stability of ranking (S_R). We construct the stability of several feature selection algorithms using normalized rank transformation and evaluate their classification accuracy using C4.5 algorithm. Therefore this work propose further understanding about the importance of stability within feature selection algorithms which can be used in deployment decision of any feature selection.

II. Similarity of Feature Selection Results

A problem with many features $f_i \in F = \{f_1, .., f_k\}$. If we assume that features are real valued, we can introduce a set of instances as V=$\{v_1, v_2, ..v_n\} \subset \Re^k$, a set of classes C and a classifier K:$\Re^k \rightarrow C$. More formally:

$$\forall\ v_i \in V,\ \ j \in \{1, .., k\}.\ v_{ij} \in f_j \qquad (1)$$

The task of feature selection algorithm is to induce a hypothesis (classifier) that accurately predicts the labels of novel instances. The learning of the classifier is inherently determined by the feature-values. Let G be some subset of F and f_G be the value vector of G. In general, the goal of feature selection can be formalized as selecting a minimum subset G such that P(C|G= f_G) is equal or as close as possible to P(C|F = f), where P(C|G = f_G) is the probability distribution of different classes given the feature values in G and P(C|F = f) is the original distribution given the feature values in F. We call such a minimum subset an *optimal* subset.

Feature is useful if it is correlated with class target; otherwise it is irrelevant. A good feature subset is one that contains feature highly correlated with or predictive of the class target. In this work we consider the following methods: Information Gain *(IG)*, Gain Ratio *(GR)*, Chi-Square *(CHI)* [2], Symmetrical uncertainty *(SYM)* [2], ReliefF *(REL)* [3] and Correlation Feature Selection (CFS) [4]. IG, GR, CHI, SYM are all feature scoring method for nominal or continuous attributes which are discretized using entropy maximization [4]. ReliefF delivers a weighting of feature while taking their interaction into account; it uses all features to compute distance among training instance.

The similarity of feature selection result particularly view in their order of rank stability against a huge amount of data managed. This stability reflects directly to the trend of over-fitting and obviously depends on how close their values are for all attributes selected. According to the measure used, feature selection algorithms produce ranking as the process of positioning features within ordinal scale in relation to every optimal subset. In order to measure similarity between

training and testing data set, we compare two ranking *r* and *r'*. We use the normalized rank transformation coefficient with Euclidean distance measure as Stability of ranking (S_R):

$$S_R(r_i, r'_i) = \sqrt{\sum_i (r_i - r'_i)^2} \quad (2)$$

where r_i *and* r'_i are the ranks of feature i in the training and testing steps respectively. An output value of SR equals to 0 means that two rankings are identical. The higher SR imply to lower stability of an algorithm.

III. Experiment Results and Discussion

We use the Coil 2000 dataset provided in UCI [5], in which there are 5900 records of training set and 4000 records of testing set. We also use 40000 records data set from PAKDD 2007[6] to examine the performance stability of feature selection algorithms. We focus on their ability to select and determine the series of order of the most relevant features with respect to a target class. In the following tables we can see the performance of these algorithms comparison.

Table 1. Dataset Coil2000 similarity and feature selection accuracy using C4.5

No.	Algo	S_R	Acc(%)	Tree	#attrib	Time
1	IG	2.10	93.89	leaves:6,size:11	**38**	**3.91**
2	SYM	2.62	93.89	leaves:6,size:11	**38**	**3.91**
3	GR	2.80	93.89	leaves:6,size:11	**38**	**3.91**
4	CHI	2.21	93.89	leaves:6,size:11	**38**	**3.91**
5	REL	1.02	93.90	leaves:6,size:11	61	7.11
6	CFS	0.43	**94.02**	leaves:2,size:2	10	0.06

Table 1 gives the stability results in SR value and classification accuracy using C4.5 classifier. The pattern of first four algorithms *(IG, SYM, GR and CHI)* gives a range SR of 2.0-3.0 that means there are different result of rank position between stage of training and testing. The common between them are accuracy, size of classification tree and duration of classification process which means, they share similar selected features (38 features). The most stable algorithm for this data set is CFS which also gives the highest and the fastest classification process.

In order to check classification characteristic, we further examine the paradigm that has been discovered in the initial data set with a bigger data set from PAKDD2007. In Table 2. we briefly can see that the first four algorithms also contribute to the similar process duration, tree size, sum of features selected and accuracy.

Moreover, we also found that the most stable algorithm is CFS, with the smallest tree structure (in size and leaves). This algorithm based on correlation between features, that ranks feature subsets according to a correlation based on heuristic evaluation function. The bias of the evaluation function is toward subsets that contain features that are highly correlated with the class and uncorrelated with each other.

Irrelevant features should be ignored because they will have low correlation with the class. Both data set also suggest that ReliefF contribute for a slightly stable in ranking the feature selection but has unreliable size of attributes selected. ReliefF algorithm also has the longest duration to classify and it is reflected by the bigger size and leaves of the classification tree. This algorithm based on feature weighting algorithm that is sensitive to feature interactions. It is likely that this result supports what we had in CFS's figures. We found that the similar features selected from four different algorithms.

Table 2. Classification result of PAKDD2007 with C4.5

No	Algo	Acc(%)	TreeSize	#attrib	Time(secs)
1	IG	**99.15**	leaves:78,size:113	**37**	**81.69**
2	SYM	**99.15**	leaves:78,size:113	**37**	**81.69**
3	GR	**99.15**	leaves:78,size:113	**37**	**81.69**
4	CHI	**99.15**	leaves:78,size:113	**37**	**81.69**
5	REL	98.60	leaves:24,size:36	29	$\gg$ 9000
6	CFS	98.82	leaves:25,size:36	39	7.09

IV. Conclusion

We defined the stability ranking measure of feature selection by deploying normalized rank transformation and classification performance. Correlation based Feature Selection (CFS) algorithm is the most stable algorithm we found on two data sets.

References

[1] I.H. Witten and E. Frank, "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[2] M.Dash and H. Liu, "Feature Selection for Classification", Intelligent Data Analysis: AAN.Int'l J.vol. 1 no.3 pp. 131-156, 1997

[3] J.Han, and M. Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann San Francisco, 2001.

[4] O. Maimon, and L.Rokach "The Data Mining and Knowledge Discovery Handbook", Springer, New York, 2005.

[5] Hettich, S. and Bay, S. D. The UCI KDD Archive [http://kdd.ics.uci.edu]. Irvine, CA: University of California, Department of Information and Computer Science. 1999.

[6] PAKDD 2007 contest.[http://lamda.nju.edu.cn/ conf/ pakdd07/dmc07/]