*2012 International Conference on Information and Network Technology (ICINT 2012)*
***IPCSIT vol. 37 (2012) © (2012) IACSIT Press, Singapore***

# Boosting Classification Accuracy Using Feature Fusion

Batnyam Nomin, Bunheang Tay and Sejong Oh [+]

Department of Nanobiomedical Science and WCU Research Center of Nanobiomedical Science, Dankook University, Cheonan 330-714, South Korea

**Abstract.** Improving classification accuracy for high dimensional dataset such as microarray is one of hot issues. Feature selection is general way to improve the accuracy by removing useless features from original dataset. In this paper, we propose a way to enhance classification accuracy using Feature Fusion Method (FFM) in addition to general feature selection schemes. Feature fusion means generating new features from combinations of original features by fusing them, i.e computing averages and multiplications of feature pairs. Proposed method is easy to understand and implement. It also can be merged with any kinds of feature selection and classification algorithms. To evaluate the performance of our method we have done experiments on 8 datasets and compared classification accuracies of original datasets and new datasets from FFM. The experiment results showed that FFM boosted the classification accuracy in many cases.

**Keywords:** Feature Selection, Classification, Feature Fusion, Accuracy

## 1. Introduction

As the size and amount of information in the world increases rapidly, a growing demand for tools and techniques that can handle them effectively has emerged. Needless to say, the bigger the dimensionality of a problem – the harder it gets to manage it efficiently. One common example is microarray data analysis. Microarray datasets generally have thousands of features; it means thousands of dimensions. By correctly classifying samples that are obtained from a patient into normal and diseased groups, it can report whether the patient has a specific disease or not. Thus the preciseness of applied classifier is extremely important. k-Nearest Neighbor (KNN) [1] and Support Vector Machine (SVM) [2] algorithms are among the best classifiers so far, which are known for their robustness and simplicity. However, their performance tends to lag behind when it comes to microarray data, where few observations with hundreds and thousands of genes (features) are given for analysis. Because most of these genes are likely to be uninformative, it often leads to a low classification accuracy and high computational cost. Thus, extraction of a feature set that holds valuable information saves time and yields better accuracy. During the recent years extensive works have been carried out on feature selection research, which proves that it is becoming a crucial part of data classification and analysis.

Feature selection methods are largely divided into three categories, filter, wrapper, and embedded approaches [3]. *Filter* methods are stand-alone techniques that are completely independent from classification algorithms that will be applied later on data sets. Its basic concept is to sequentially search and evaluate features and select good subset using simple statistics. *Wrapper* models incorporate the classifier that will be used to evaluate the features. In the *embedded* technique, the search for an optimal subset of features is built into the classifier construction, and can be seen as a search in the combined space of feature subsets and hypotheses. Just like wrapper approaches, embedded approaches are thus specific to a given learning algorithm. In our experiment we consider RFS [4], a distance discriminant method FSDD [5], and ReliefF [6]. RFS is based on a dataset evaluation measurer R-value [7].

[+] Corresponding author. Tel.: +82-41-550-3484; fax: +82-41-550-1149.
*E-mail address*: sejongoh@dankook.ac.kr.

In this paper we propose Feature Fusion Method (FFM) to improve classification approach for high dimensional datasets. Our idea is to generate new dataset from original dataset through fusion of features. The initial idea was introduced in [8]. They evaluate each feature using statistical measure and merge highly ranked features. In contrast, we adopt measures in feature selection algorithms. For the fusion method, we test average (*avg*) and multiplication (*mult*) value of pair of features. Detail is described in the next section.

This paper is organized as follows: Section 2 covers the idea of Feature Fusion method, also datasets and parameters that were used to conduct the experiment. Section 3 shows experiment results, and finally, Section 4 discusses our future research and concludes this work.

# 2. Materials and Methods

## 2.1. Feature Fusion Method

Before we describe the proposed approach we need to understand Feature Fusion. For simplicity, we consider fusion of two features. Let's suppose $f_i$ and $f_j$ are two features from $n$-dimensional dataset with $m$ instances. Feature can be expressed by set of feature values as follows:

$$f_i = \{x_{i1}, x_{i2}, x_{i3}, \ldots, x_{im}\}, f_j = \{x_{j1}, x_{j2}, x_{j3}, \ldots, x_{jm}\} \quad (1)$$

By fusing $f_i$ and $f_j$, we can get a new feature $f_k$. We test two fusion methods, average and multiplication.

$$(avg): \quad f_k = \{\frac{x_{i1}+x_{j1}}{2}, \frac{x_{i2}+x_{j2}}{2}, \frac{x_{i2}+x_{j2}}{2}, \ldots, \frac{x_{im}+x_{jm}}{2}\} \quad (2)$$

$$(mult): \quad f_k = \{(x_{i1}\times x_{j1}), (x_{i2}\times x_{j2}), (x_{i3}\times x_{j3}), \ldots, (x_{im}\times x_{jm})\} \quad (3)$$

The steps of FFM is depicted on Figure 1 and explained in details below:

1. Prepare an original dataset *OD*.
2. Apply feature selection on *OD* and select the best 40 features.
3. Apply Feature Fusion on *OD*' with 40 features and take two new datasets $ND_{avg}$ and $ND_{mult}$.
   The original 40 features are also included in $ND_{avg}$ and $ND_{mult}$, and the number of features of $ND_{avg}$ and $ND_{mult}$ is ${}_{40}C_2 + 40$.
4. Apply feature selection and do classification test.
   New datasets $ND_{avg}$ and $ND_{mult}$ have numerous features and feature selection is required before classification test.

We can choose various algorithms for classification test and feature selection. In other word, FFM is independent from specific classification and feature selection schemes.

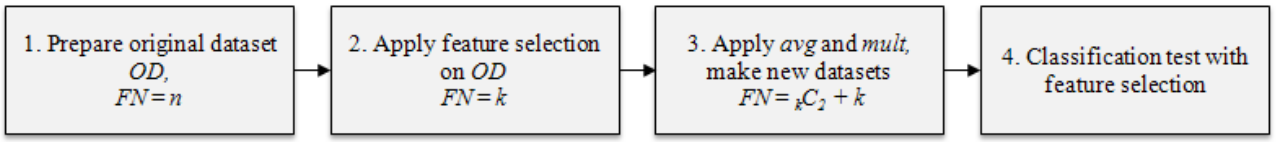


Fig. 1: Generic steps of FFM. (*FN*: feature number of dataset)

## 2.2. Datasets and Parameters

To test the effectiveness of FF we carried out experiment on eight datasets, four of which are from UCI Machine Learning repository [9], and the other half is microarray. The properties of datasets are summarized in Table 1. To avoid overfitting problem, we adopt *k*-fold validation. For datasets with few samples we used 2-fold validation, for Sonar dataset *k* is 3, and for Multiple Features and Madelon datasets which have 2000 instances we choose *k* as 5. Two popular classifiers are considered to estimate the performance of Future Fusion method. For feature selection, we adopt RFS, FSDD, and ReliefF. Parameter values for classification and feature selection algorithm are presented in Table 2 and 3; 7 nearest neighbours and Euclidian distance

were used for KNN classifier and linear kernel for SVM. To get the best accuracy from given dataset, we test different number of features, 5, 10, 15, and so on.

Table. 1: Dataset list

| No | Data Set Name | Features | Classes | Instances | Folds | Ref |
|---|---|---|---|---|---|---|
| 1 | Prostate | 12600 | 2 | 102 | 2 | [10] |
| 2 | Arcene | 10000 | 2 | 100 | 2 | [9] |
| 3 | Duke | 7129 | 2 | 44 | 2 | [12] |
| 4 | BrcaEr | 754 | 2 | 146 | 2 | [11] |
| 5 | DLBCL | 661 | 3 | 141 | 2 | [11] |
| 6 | Multi Features | 649 | 10 | 2000 | 5 | [9] |
| 7 | Madelon | 500 | 2 | 2000 | 5 | [9] |
| 8 | Sonar | 60 | 2 | 416 | 3 | [9] |

Table. 2: Parameters related with feature selection algorithm

| Algorithm | Parameters |
|---|---|
| RFS | $K = 7,\ \theta = 4$ |
| FSDD | $\beta = 1$ |
| ReliefF | $K = 10$, *repeat time* = (the number of instances)/2 |

Table. 3: Testing condition related with classifiers

| Classifier | Testing condition |
|---|---|
| KNN | $K = 7$, distance function = simple Euclidean distance |
| SVM | Kernel = linear kernel |

## 3. Result

Figure 2 presents result of classification analysis. We compared original (*orig*) dataset, on which we employ only conventional feature selection and classifier, and new *avg* and *mult* datasets. From the graphs it's obvious that Feature Fusion methods (*avg* and *mult*) improve accuracy in many cases. In some cases *avg* and *mult* produce much higher accuracy than the original. Especially this performance can be observed from ReliefF results. For example, the improvement in *KNN+ReliefF+FFM* for the Arcene dataset jumped from *orig* 0.11% to 0.71%, and in the BrcaEr dataset, it increased even from 0.027% to 0.89%. *ReliefF+FFM* approach has delivered a decreasing accuracy only in 3 cases out of 16. Hence it is certain that the improvement is not only substantial, but significant. In the half of the experiments we got slight increase for *RFS+FFM* and *FSDD+FFM*. Classification for *orig*, *avg* and *mult* was performed 3 times for KNN, and 3 times for SVM for one dataset. Thus 6 times for every dataset and in total we performed 6x8=48 classifications for each of *orig*, *avg*, and *mult*. Result on Figure 2 shows that for FFM *avg* 26 cases from 48 are improved, and for FFM *mult* 23 from 48, i.e 50% of the total experiment delivered a better result. For each dataset we selected highest original accuracy and found an improved version among its corresponding *avg* and *mult* results. Then in order to get the accuracy improvement (Table 4) we computed their differentiation.

## 4. Discussion and Conclusion

In this study we have presented a simple but efficient way to boost a classification accuracy. The idea of this approach is to choose a good feature set from the original data using popular algorithms and produce "artificial" features through FFM technique. Then again better features are selected from the obtained artificial feature set. For the proposed method traditional feature selector plays important role, since in the

first stage it selects valuable feature set. Thus it is essential to choose an optimal feature selection. FFM is a completely independent approach, so it can co-operate with other feature selection methods. Moreover, FFM can be calculated not only by *avg* and *mult*, but also in different calculations, for instance, *sum*, *diff*, or even a combination of more than two features. In the future, we will further work on the improvement of FFM by experimenting various ways of generating informative features.

Table. 4: Summary of accuracy improvement (%)

| No | Data Set Name | Best improvement | Used FS | Used FFM | Used Classif. |
|---|---|---|---|---|---|
| 1 | Prostate | 4.9 | FSDD | *mult* | KNN |
| 2 | Arcene | 7.0 | ReliefF | *avg, mult* | KNN |
| 3 | Duke | 11.9 | RFS | *avg* | KNN |
| 4 | BrcaEr | 1.4 | RFS | *mult* | SVM |
| 5 | DLBCL | 0.7 | RFS | *mult* | KNN |
| 6 | Multi Features | 20.7 | RFS | *avg* | SVM |
| 7 | Madelon | 20.8 | FSDD | *avg* | KNN |
| 8 | Sonar | 1.9 | RFS | *avg* | KNN |

## 5. Acknowledgements

This study was supported by grant No. R31-2008-000-10069-0 from the World Class University (WCU) project of the Ministry of Education, Science & Technology (MEST) and the Korea Science and Engineering Foundation (KOSEF).

## 6. References

[1] T. Cover, and P. Hart. Nearest neighbor pattern classification, *IEEE Transactions*. 1967, **13**(1): 21-27

[2] CC. Chang, and CJ. Lin, LIBSVM: a library for support vector machines, *Science*. 2001, **2**:1-39.

[3] I. Guyon, and A. Elisseeff. An introduction to variable and feature selection. *J of Mach Learn Res*. 2003, **3**: 1157-1182

[4] L. Lee, et al. RFS: efficient feature selection method based on R-value, *Computers in Biology and Medicine*, submitted for publication.

[5] J. Liang, et al. Invariant optimal feature selection: A distance discriminant and feature ranking based solution, Pattern *Recognition*. 2008, **41**(5): 1429-1439.

[6] Y. Sun and D. Wu. A RELIEF based feature extraction algorithm, *Proceedings of the 2008 SIAM International Conference on Data Mining*. 2008, 188-195.

[7] S. Oh, A new dataset evaluation method based on category overlap, *Computers in Biology and Medicine.* 2011, **41**(2): 115-122.

[8] P. Chopra et al. Improving cancer classification accuracy using gene pairs, *PloS One*. 2010, **5**(12):14305.

[9] UCI Machine Learning Repository. http://archive.ics.uci.edu/ml.

[10] D. Singh et al. Gene expression correlates of clinical prostate cancer behavior, *Cancer Cell*. 2002, **1**(2): 203-209.

[11] Y. Hoshida et al. Subclass mapping: identifying common subtypes in independent disease data sets. *PLoS One.* 2007, **2**(11):e1195.

[12] M. West, et al. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences*. 2001, 11462-11467.

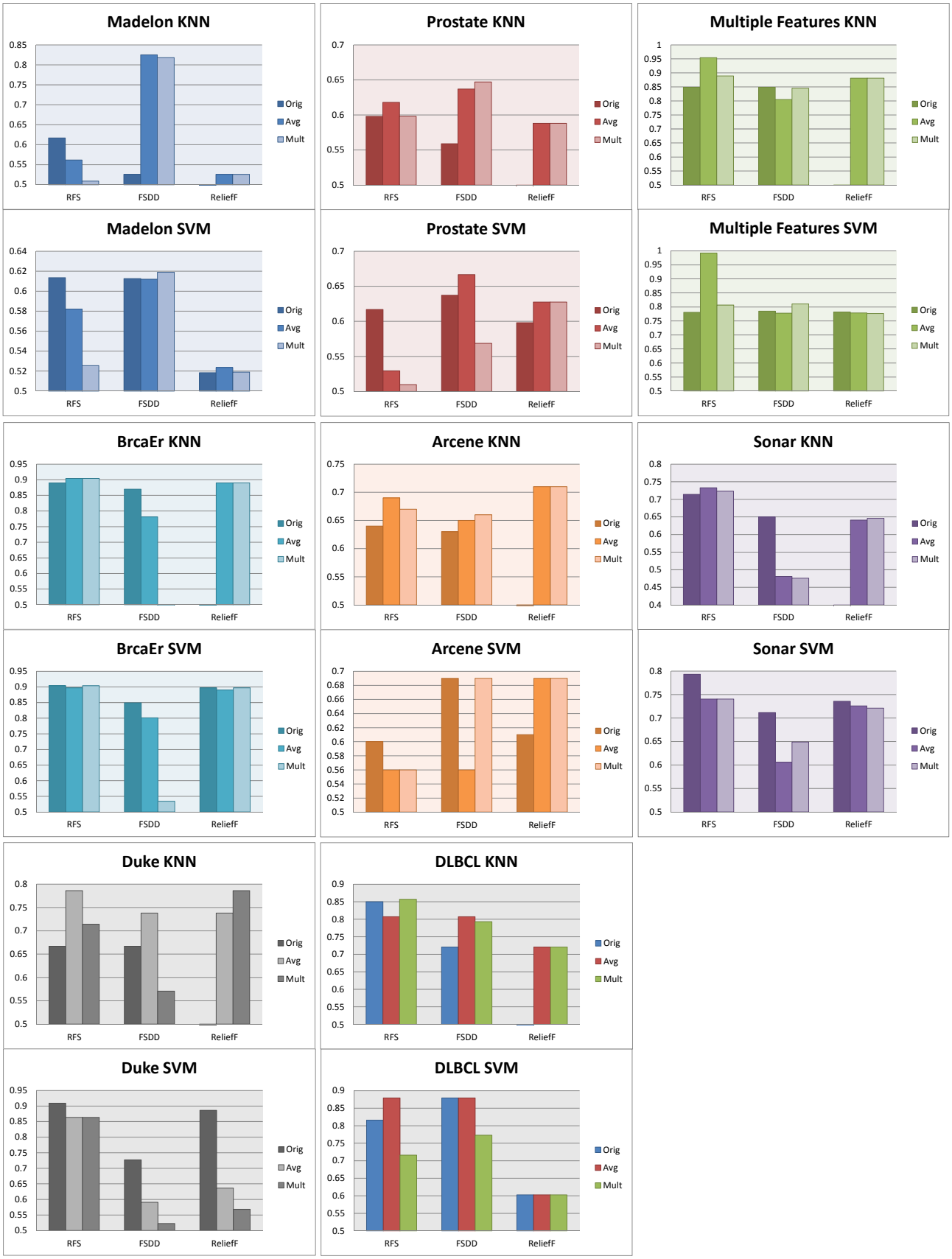


Fig. 2: Classification accuracy of Original, *avg*, and *mult* set of features.