An Efficient Classification for Single Nucleotide Polymorphism (SNP) Dataset

Nomin Batnyam, Ariundelger Gantulga, and Sejong Oh*

Abstract. Recently, a Single Nucleotide Polymorphism (SNP) which is a unit of genetic variations has caught much attention as it is associated with complex diseases. Various machine learning techniques have been applied on SNP data to distinguish human individuals affected with diseases from healthy ones or predict their predisposition. However, due to its data format and enormous feature space SNP analysis is a complicated task. In this research an efficient method is proposed to facilitate the SNP data classification. The aim was to find the most effective way of SNP data analysis by combining various existing techniques. The experiment was conducted on four SNP datasets obtained from the NCBI Gene Expression Omnibus (GEO) website, two of them are from patients with mental disorders and their healthy parents; and the other two are cancer related data. The analysis process consists of three stages: first, reduction of feature space and selection of informative SNPs; next, generation of an artificial feature from the selects SNPs; and last but not least, classification and validation. The proposed approach proved to be effective by distinguishing two groups of individuals with high accuracy, sometimes even reaching 100% preciseness.

Keywords: Single Nucleotide Polymorphism (SNP), classification, feature selection.

1 Introduction

Human genome consists of approximately three billion DNA base pairs, called nucleotide. Nearly 99% of them are identical among all humans (population), and

Nomin Batnyam · Ariundelger Gantulga · Sejong Oh
Department of Nanobiomedical Science and WCU Research Center, Dankook University, Cheonan 330-714, South Korea
e-mail: {gngrfish,ariuka_family}@yahoo.com
dkumango@gmail.com

* Corresponding author.

R. Lee (Ed.): *Computer and Information Science*, SCI 493, pp. 171–185.
DOI: 10.1007/978-3-319-00804-2_13

only one percent varies among individuals. A large portion of these genetic variations occur as Single Nucleotide Polymorphisms (SNPs). Studies have shown that SNPs may have important biological effects, such as association with complex diseases and different reactions to medications and treatments. Also, it has several advantages over microarray gene expressions, such as it is unlikely to change over time. That is, SNPs of a patient at a birth will remain same whole life. It is much easier and faster to collect SNP sample, for it can be obtained from any tissues in the body, while microarray sample must be taken only from specific tissues [1]. Consequently, gene mapping and detection of polymorphisms have caught much attention recently, and currently enormous number of genetic variations is being discovered and analyzed. Many machine learning techniques have been proposed and applied to SNP data classification. However, it is facing several challenges due to its high volume, which poses computational time complexity and low accuracy.

There is no universally optimal method that fits well with every type of data. Therefore, the aim of this study is to find an efficient way of SNP data analysis, by combining known approaches such as feature selection, R-value evaluation, feature fusion and classification to achieve higher accuracy and less time consumption. In our experiment we utilized two powerful and well-known classifiers k-Nearest Neighbor and Support Vector Machines; and an Artificial Gene Making classifier. However, because the number of samples in a SNP dataset is undue small relative to its attribute size, the curse of dimension occurs. To overcome this problem we reduce a feature space through feature selection. The feature selection plays a crucial role in classification of SNP, since it not only solves the problem of dimension, but also lowers time complexity and facilitates the accuracy improvement by selecting the most informative set of attributes. To perform this task we employed four algorithms, two of which are popular methods, Feature Selection based on Distance Discriminant and ReliefF; and two methods, R-value based Feature Selection and an Algorithm based on Feature Clearness.

The remainder of this paper is organized as follows. In section 2, brief introductions to machine learning techniques that were applied in this study are described. We will provide background information of feature selection and classification algorithms. Datasets and proposed method are described in Section 3. In Section 4 we present experimental results and conclude this work in Section 5.

2 Related Work

2.1 Feature Selection

Like microarray gene expression, SNP data has high dimensionality, but several hundred times bigger, making the data analysis impractical. The whole dataset is

composed of informative polymorphisms that are often called Tag SNPs, as well as irrelevant ones. Thus elimination of useless SNPs and extraction of a small subset of discriminative ones will help identify tag SNPs that can be used as biomarkers or other cause associated polymorphisms. This process can be done through feature selection, which also facilitates classification task by reducing a search space.

Feature selection is largely divided into filter and wrapper types. Filter methods are applied on a dataset before classification task and usually evaluate features based on simple statistics such as *t*- or *F*-statistics, or *p*-value [3]. Contrarily, wrapper methods make use of learning algorithm to select a feature subset by incorporating it inside the feature search and selection. The latter approach has an advantage over filter method of showing better performance for particular learning algorithms. However, it is more computationally expensive [4]. In our experiments we used only filter methods. Two popular approaches are: a Feature Selection based on Distance Discriminant (FSDD) [5] and a feature weight based ReliefF [6]; and R-value based Feature Selection (RFS) [7] and an Algorithm based on Feature Clearness (CBFS) [8].

2.1.1 ReliefF

ReliefF algorithms are able to detect conditional dependencies between features. The original Relief [2] algorithm can be used to select nominal and numerical attributes, but it is limited to binary class problems. Meanwhile, the newer and more robust version ReliefF is for multiple class problems and capable of dealing with incomplete and noisy data [6]. Main drawback of Relief algorithms is its time complexity compared to other methods in the literature.

In general, Relief algorithm produces quality estimation for each attribute. To do so, it searches two nearest neighbors for a random sample: one from the same class and one from a different class, and updates quality estimation for each feature depending on the values of the sample and its nearest neighbors.

2.1.2 Feature Selection Based on Distance Discriminant (FSDD)

The main advantage of this algorithm is that it produces as optimal result as exhaustive search methods, in the meantime, it makes use of a feature ranking scheme to solve the problem of computational complexity. The basis of FSDD is to find features with good class separability among different classes as well as make samples in the same class as close to one another as possible. A criterion for feature is a difference between the distance within classes d_w multiplied by a user defined value β, which works as an impact controller and usually set to 2, and the distance between different classes d_b, see Equation (1) [5].

$$d_b - \beta d_w \tag{1}$$

2.1.3 Feature Selection Based on R-value (RFS)

RFS is a simple feature ranking algorithm based on a dataset evaluation measure R-value [18]. R-value measures the quality of a dataset assuming that the separability of classes is strongly related to category overlap. It captures overlapping areas among classes in a dataset, the lower the value of R the more separable are the classes from one another. Accordingly, RFS algorithm scores the overlapping areas of classes for each feature without considering relationship among features. RFS calculates the R value for a feature F_i by dividing the total number of feature values in a target feature that belongs to overlapping area by total number of samples of given dataset, see Equation (2) [7].

$R(F_i)$ = (total number of feature values in F_i that belongs to overlapping areas)/(total number of samples of given dataset) (2)

2.1.4 Algorithm Based on Feature Clearness (CBFS)

Likewise the above mentioned algorithms, CBFS measures separability of classes in attributes. It adopts CScore which estimates the degree of correctly clustered samples to the centroid of their class. Its approach works by calculating distance between the target sample and centroid of each class, then compares the class of the nearest centroid with the class of the target samples. Classification accuracy of training data serves as clearness value for a feature [8]. One of CBFS strong points is that it can be combined with other feature ranking algorithms; in our experiment we merged it with the R-value.

2.2 Classifiers

Classification is a class predicting process of a data analysis based on supervised learning method. In order to find a class for unknown sample it uses information of samples that already belong to specific groups. Numerous supervised learning algorithms have been developed, in this study we employed two widely used machine learning techniques, k-Nearest Neighbor (KNN) [9] and Support Vector Machine (SVM) [10]; and an Artificial Gene Making [11] method.

2.2.1 K-Nearest Neighbor (KNN)

KNN is one of the most popular machine learning algorithms and is famous for its simplicity and effectiveness. Also it can be applied in a variety of fields. The algorithm classifies objects based on closest training examples in the feature space and by majority vote of its neighbors. Usually Euclidian distance is used as the distance metric for continuous variables. For a target sample KNN algorithm finds the k closest samples in the learning set and predicts its class by assigning the

target sample to the class whose samples are the most common among those k neighbors [3].

2.2.2 Support Vector Machines (SVM)

SVM is a non-probabilistic binary linear classifier based on regularization techniques. To do a classification task, it constructs a hyperplane in a feature space that serves as a boundary between two class samples. The larger the distance between the hyperplane and the nearest sample, the better the classification result. The boundary is determined by the probability distribution or based on the classification of the training patterns. When SVM is applied on a multiclass dataset, the problem can be decomposed into a set of binary problems, and then combined to make a final multiclass prediction [19].

2.2.3 Artificial Gene Making (AGM/Alpha) Method

The Artificial Gene Making method, or sometimes referred to as Alpha, was initially devised for classification of microarray dataset. It is confirmed that a congestion area among classes is one of main reasons of low classification result [18]. The main purpose of Alpha is to reduce this congestion area by adding an artificial gene $(n + 1)$ to a dataset with n features, assuming that, for example, a dataset with two dimensions (genes) may become easier to classify if a third dimension is added. First, the dataset is divided into training and test. For construction of a new gene, there is a need to determine two important values α and β, where the first one measures the quality of the original dataset and it expresses a matching ratio between the predicted and original class labels of the training data. The β is dependent on α and estimates the congestion area. The values of a new gene in a training data is calculated by (*original class label*) x β-value, and the gene for test data is calculated by (*predicted class label*) x β-value. After the production of a new gene or feature is done, the classification procedure can be performed [11].

2.3 Feature Fusion Method (FFM)

Feature Fusion (FFM) [20] method is simple, yet proved to be effective in many cases when applied on microarray and other biological datasets. It was confirmed that when attributes or genes are combined with other attributes or genes, they frequently improve the classification accuracy thanks to the mutual information and interaction of features [21]. Therefore, we assume that the same technique could be successfully used on SNP dataset.

The idea of the approach is to produce a new dataset by fusing features of the original one, and the fusion can be done in number of ways. That is, values of one feature can be multiplied, averaged, subtracted, or added up with values of another feature. In addition, the number of features to be fused can range from two to as many as an experimenter ventures to try. For example, in case of two features f_i

and f_j with feature values x_n in Equation (3) a new feature constructed by FFM will be calculated as shown in Equation (4).

$$f_i = \{x_{i1}, x_{i2}, x_{i3}, \ldots, x_{in}\} \\ f_j = \{x_{j1}, x_{j2}, x_{j3}, \ldots, x_{jn}\} \tag{3}$$

$$\text{FFM } (avg): f_k = \{(x_{i1} + x_{j1})/2, (x_{i2} + x_{j2})/2, (x_{i2} + x_{j2})/2, \ldots, (x_{in} + x_{jn})/2\} \\ \text{FFM } (mult): f_k = \{(x_{i1} \text{ x } x_{j1}), (x_{i2} \text{ x } x_{j2}), (x_{i2} \text{ x } x_{j2}), \ldots, (x_{in} \text{ x } x_{jn})\} \tag{4}$$

3 Methods and Datasets

3.1 Datasets

The SNP datasets tested in this experiment were downloaded from the NCBI Gene Expression Omnibus (GEO) repository [12]. GEO is a public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomic data.

Our experiment was performed on four Affymetrix Mapping 250K Nsp SNP Arrays: GSE9222 [13], GSE13117 [14], GSE16125 [15], and GSE16619 [16], where the first two are related to mental disorders and the latter two are associated with cancers. Dataset features consist of over 250,000 SNPs and two class labels (case and control). GSE16125's SNP values are in the form of continuous numbers, but the other three arrays have alphabetical format. That is, each sample can have one of four SNP markers: AA, AB, BB, and No Call. AA and BB represent homozygous genotype, the AB represents heterozygous genotype, and No Call is a missing value. Human body contains two copies of each gene, one from father and one from mother. If a mutation occurs in one copy of the gene, it is called heterozygous genotype. However, if both copies of a gene are mutated then that individual is considered homozygous. In order to apply feature selection and classification to these three datasets, the alphabetical format has to be transformed into numerical. There are several ways to do so, but in our case, we adopted a method of three binary values [17] for each genotype shown in Equation (5).

$$AA = 100, \quad B = 010, \quad AB = 001, \quad No\ Call = 000 \tag{5}$$

Here we present two datasets results and other two are included in supplementary material. All four datasets summary can be found in Table 1 and brief description of two datasets which included in this paper is below.

GSE13117 series is composed of 120 cases with unexplained mental retardation along with their healthy parents as control.

GSE16619 series is a breast cancer dataset composed of 105 control and 159 case samples; and approximately 500,000 SNPs.

Table 1 Summary of SNP datasets

No	Dataset name	# of SNPs	# of Samples	Case information	Ref.
1	GSE9222	250,000+	567	Autism (ASD)	[13]
2	GSE13117	250,000+	432	Mental Retardation	[14]
3	GSE16125	250,000+	48	Colon Cancer	[15]
4	GSE16619	500,000+	111	Breast Cancer	[16]

3.2 Methods

Our experiment has three main steps. First of all, a feature selection is performed on the whole data to pick informative set of SNPs and reduce the feature space. Each dataset has undergone a feature selection procedure by algorithms described in previous sections and the top 10 to 100 SNPs were chosen. On the second stage a new dataset is produced from the feature selected data. The details of this step are disclosed below, and the last but not least step is classification. The steps were validated by classification algorithms using 10-fold cross-validation. For KNN we consider the closest 7 neighbors ($k = 7$) because this value was found to produce the best accuracy in most cases. We test SVM with kernel such as linear, polynomial, RBF (radial basis function), and sigmoid. When selecting the linear kernel, we get the best accuracy.

3.2.1 Feature Fusion Method (FFM)

In this study we adopted multiplication (*mult*) and average (*avg*) methods of FFM to experiment on SNP data, i.e. create a new feature from the existing feature pairs by multiplying or averaging feature values for all samples.

The artificial data were produced in two different ways: through Original FFM and R-value FFM. Original FFM means generating new features by multiplying or averaging the combination of the original n features and without performing any additional techniques (Fig. 1). The number of generated SNPs in a new dataset equals ${}_nC_2$. The R-value FFM, on the other hand, generates new features only from features that are evaluated and then selected by R-value measures. The latter approach is described in the following subsection.

3.2.2 R-value Evaluation

When generating new attributes from the original ones, the feature space grows dramatically. For example, if we create a combination of feature pairs from a subset of ten (${}_{10}C_2$), then the number of newly generated features will be 45, which is over four times bigger than the initial size. Therefore, while producing artificial SNPs, choosing only good ones for further analysis is required. R-value (*Rval*)

evaluation is, sort of, performing the role of feature selection during the FFM process. Thus the number of SNPs in the new dataset is random depending on the quality of feature pairs.

All features' R-value is calculated and smaller value indicates a better feature. For example, for features x and y, calculated measure will be $Rval(x)$ and $Rval(y)$ respectively. Then R-value is calculated for every combination of two features constructed by *mult* or *avg* methods $Rval(mult(x, y))$ or $Rval(avg(x, y))$. If the value of a new feature is less than the value of the original single feature, then the new feature combination is taken, otherwise the original is taken. Generic steps for R-value FFM are depicted on Fig. 2.

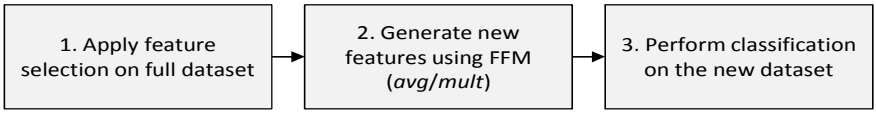

Fig. 1 Generic steps of Original FFM

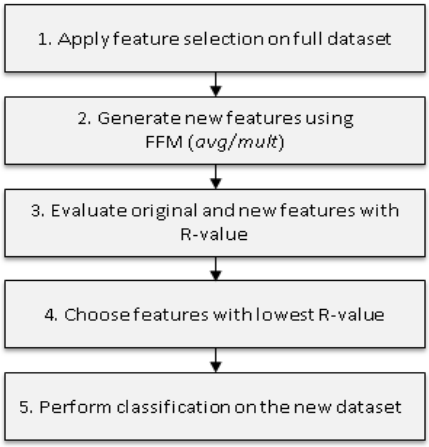

Fig. 2 Generic steps of R-value FFM

4 Results

This section reviews experimental results for datasets: GSE16619 and GSE13117. We will go through datasets one by one. One dataset has three detailed classification accuracy tables - one for each classifier (Supplementary Tables S7 to S18). First column is a feature selection, followed by a FF method, and the last column shows the average accuracy of top 100 SNPs. The top row of accuracy results is Original data, and below are accuracies of new datasets generated by FF methods. The results are compared to the original, and if the accuracy of a new dataset is improved or remained equal, it is marked bold.

4.1 GSE16619 Series

The average classification accuracy for GSE16619 series is summarized in Table 4. From the Table 4 we can see that the average accuracy of the entire new data constructed by FSDD+FFM and classified by Alpha; constructed by RFS + FFM and classified by KNN; constructed by ReliefF + FFM and RFS + FFM and classified by SVM is improved. However, if compare the overall accuracies of FF methods in Table 3 and Fig. 5, then again, the approaches are performing at around similar level, but as for feature selection Table 2 and Fig. 6 show that the CBFS is the best.

Table 2 Comparison of Feature Selection's overall classification accuracy for GSE16619 series. The best results are marked bold.

GSE16619	ReliefF	RFS	FSDD	CBFS
Alpha	0.486667	0.457333	0.486833	**0.665**
KNN	0.51	0.507167	0.565667	**0.684**
SVM	0.580483	0.51285	0.56445	**0.934667**

Table 3 Comparison of Feature Fusion's overall classification accuracy for GSE16619 series. The best improvements are marked bold.

GSE16619	Original	Multiply	Average	Mult+R	Avg+R
Alpha	0.55375	0.482	0.557	0.49325	**0.57425**
KNN	0.57225	0.54775	0.5835	0.55775	**0.59025**
SVM	0.6406	**0.6575**	0.6456	0.64305	0.65075

Table 4 Summary of average classification accuracies of top 100 SNPs for GSE16619 series

GSE16619		Original	Multiply	Average	Mult+R	Avg+R
	ReliefF	0.597	0.411	0.546	0.409	0.545
Alpha	RFS	0.508	0.423	0.47	0.443	0.476
	FSDD	0.473	**0.487**	**0.478**	**0.514**	**0.478**
	CBFS	0.637	0.607	**0.734**	0.607	**0.798**
	ReliefF	0.531	0.481	0.526	0.497	**0.542**
KNN	RFS	0.49	**0.502**	**0.514**	**0.523**	**0.511**
	FSDD	0.561	**0.57**	**0.572**	**0.569**	0.552
	CBFS	0.707	0.638	**0.722**	0.642	**0.756**
	ReliefF	0.5589	**0.5958**	**0.5652**	**0.6057**	**0.5787**
SVM	RFS	0.4941	0.5364	0.5058	0.5058	0.5139
	FSDD	0.5769	0.5535	0.567	0.5598	0.5688
	CBFS	0.9325	**0.9443**	**0.9444**	0.9009	**0.9416**

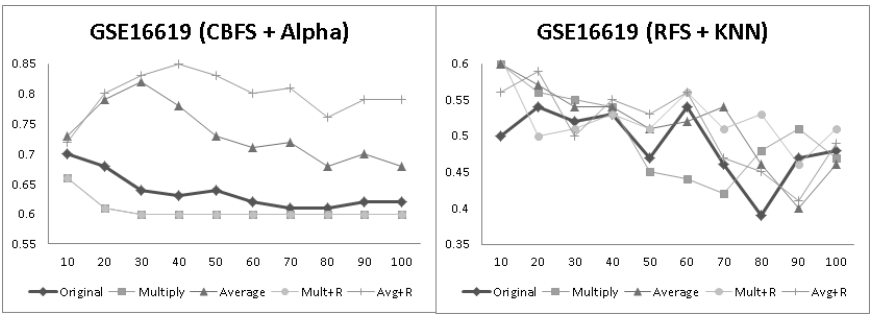

Fig. 3 Accuracy for GSE16619, a) feature selected by CBFS and classified by Alpha (on left); b) feature selected by RFS and classified by KNN (on right) and classified by Alpha. Thick line depicts the original data and other colors are representing new data constructed by FFM, the y and x axes indicate accuracy and number of SNPs respectively.

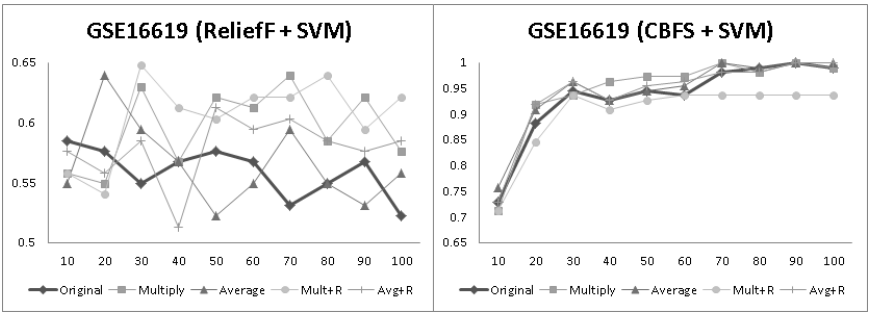

Fig. 4 Accuracy for GSE16619, a) feature selected by ReliefF (on left) and b) CBFS (on right) and classified by SVM. Thick line depicts the original data and other colors are representing new data constructed by FFM, the y and x axes indicate accuracy and number of SNPs respectively.

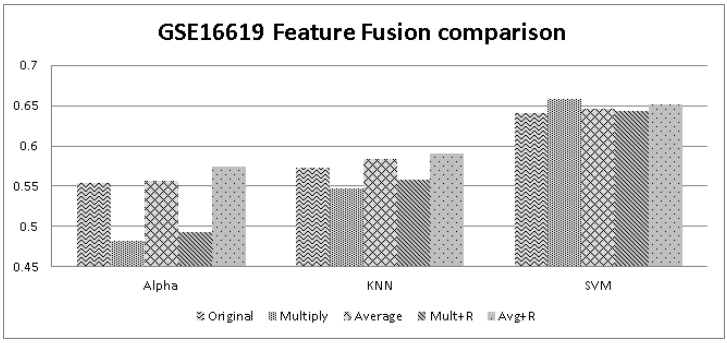

Fig. 5 Comparison chart of Feature Fusion methods using the overall accuracy of top 100 SNPs from GSE16619 series

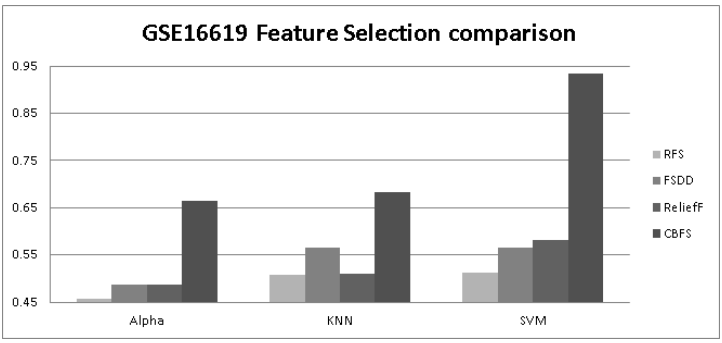

Fig. 6 Comparison chart of Feature Selection using the overall accuracy of top 100 SNPs from GSE16619 series

4.2 GSE13117 series

Due to a high time complexity, we consider only the top 50 original SNPs for generating a new data by R-value FFM. However, from Fig. 7 and Fig. 8 we can see that the said datasets (Mult+R and Avg+R) have a high potential for further accuracy improvement. In addition, Table 5 shows in general that over 4/5 of the newly generated data gave better results than the original.

As for feature selection, CBFS is still performing better than the others, but not outdoing so much compared to other datasets (Fig. 9). Also from Fig. 9 and Fig. 10 we can assume that SVM classifies GSE13117 dataset the best, reaching 99% accuracy (Fig. 8).

Table 5 Summary of average classification accuracies for GSE13117 series

GSE13117		Original	Multiply	Average	Mult+R	Avg+R
	ReliefF	0.6234	**0.6331**	0.614	0.5584	0.549
Alpha	RFS	0.3401	**0.3528**	**0.3497**	**0.353**	**0.352**
	FSDD	0.4555	**0.5288**	**0.4947**	**0.509**	**0.4694**
	CBFS	0.667	**0.6672**	**0.671**	**0.6674**	**0.673**
	ReliefF	0.6042	**0.6217**	**0.6045**	**0.6596**	**0.6212**
KNN	RFS	0.5715	**0.5793**	**0.5877**	**0.6282**	**0.602**
	FSDD	0.5863	**0.6382**	**0.6377**	**0.6076**	**0.6048**
	CBFS	0.691	0.6875	**0.6932**	**0.7112**	**0.7286**
	ReliefF	0.6426	**0.9107**	0.6408	**0.86**	**0.6494**
SVM	RFS	0.6404	**0.8664**	0.6396	**0.7346**	**0.6662**
	FSDD	0.7236	**0.9002**	**0.7381**	**0.7832**	0.673
	CBFS	0.8579	**0.9184**	**0.8903**	0.7906	0.8222

Table 6 Comparison of Feature Selection's overall classification accuracy for GSE13117 series. The best results are marked bold.

GSE13117	ReliefF	RFS	FSDD	CBFS
Alpha	0.602183	0.349567	0.4977	**0.6688**
KNN	0.62215	0.591383	0.6188	**0.69965**
SVM	0.767633	0.73585	0.7856	**0.8661**

Table 7 Comparison of Feature Fusion's overall classification accuracy for GSE13117 series. The best improvements are marked bold.

GSE13117	Original	Multiply	Average	Mult+R	Avg+R
Alpha	0.5215	**0.545475**	0.53235	0.52195	0.51085
KNN	0.61325	0.631675	0.630775	**0.65165**	0.63915
SVM	0.71612	**0.898925**	0.7272	0.7921	0.7027

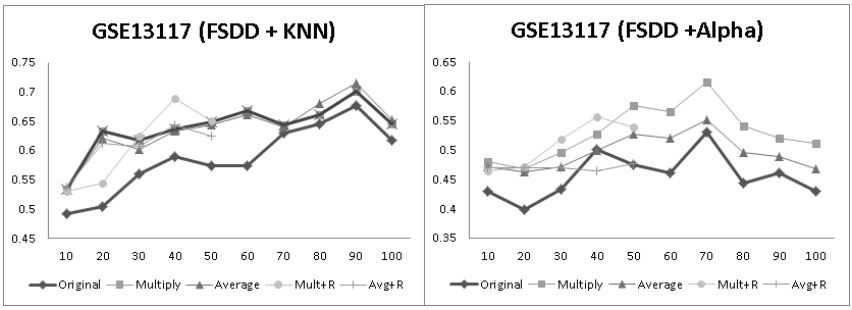

Fig. 7 Accuracy for GSE13117 a) feature selected by FSDD and classified by KNN (on left); b) feature selected by FSDD and classified by Alpha (on right). Thick blue line depicts the original data and other colors are representing new data constructed by FFM, the y and x axes indicate accuracy and number of SNPs respectively.

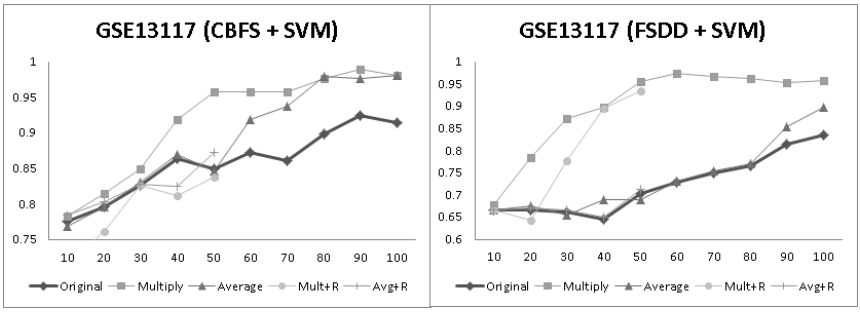

Fig. 8 Accuracy for GSE13117, a) feature selected by CBFS (on left); b) feature selected by FSDD (on right) and classified by SVM. Thick blue line depicts the original data and other colors are representing new data constructed by FFM, the y and x axes indicate accuracy and number of SNPs respectively.

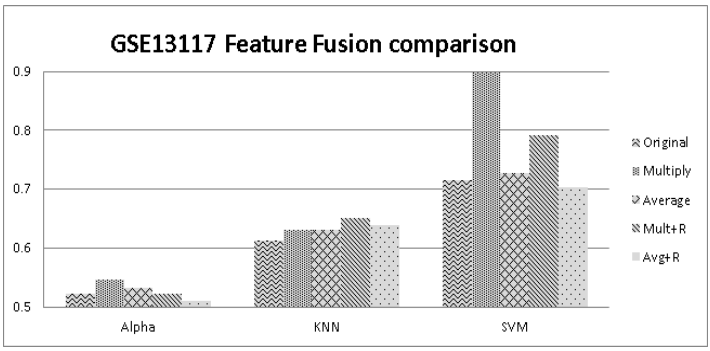

Fig. 9 Comparison chart of Feature Fusion methods using the overall accuracy of top 100 SNPs from GSE13117 series

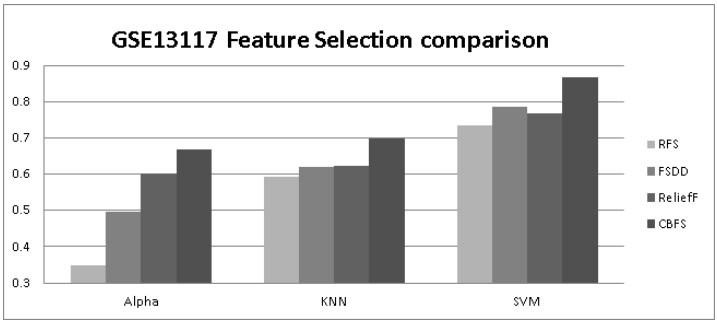

Fig. 10 Comparison chart of Feature Selection using the overall accuracy of top 100 SNPs from GSE13117 series

5 Discussion and Conclusion

Studies have shown that genetic variations are associated with various diseases and they play an important role in the determination of individual's susceptibility to those diseases. SNP is the most common genetic variation, thus machine learning techniques are increasingly applied to identify the interaction between SNPs and complex diseases. However, reaching high classification accuracy for such type of data is not an easy task, since SNP is composed of only four values.

The aim of this research is to suggest an as effective approach as possible to enhance a SNP dataset classification accuracy, and we believe that the goal was attained improving more than the half of the results. The experiment was conducted on four SNP data series that are related to mental disorders and cancers. To predict the diseases we employed classifiers, feature selections, as well as feature evaluation and generation techniques. Some algorithms work well on some data, but may show average performance on others. In general, the approach is composed of three main steps, which are 1) a selection of the most relevant SNPs, 2) a generation of new SNPs from the selected ones, and 3) a classification using a 10-fold cross validation.

If we compare our results with the previous studies [17], which is also experimented on the same SNP data sets (GSE9222 and GSE13117), we got significantly better accuracies. In case of GSE13117, one of our new datasets selected and generated by CBFS + FFM could achieve a 99% accuracy by SVM classifier (Supplementary Table S15), while the previous studies' highest could reach only 66%. From the Feature Fusion comparison tables and charts (Tables 3,7 and Fig. 5,10) we can see the original FFM and R-value FFM perform at around same level, however, from the Feature Selection comparison tables and charts (Tables 2,6 and Fig. 6,9) it is clear that CBFS algorithms is far better for all datasets. The best accuracy of GSE16125 is 82.5%, GSE16619 is 100%, GSE13117 is 99%, and GSE9222 is 78.8% (refer to Supplementary material).

Therefore, from these results we conclude that SNPs can be effectively used to distinguish individuals with complex diseases from the healthy ones, and the proposed approach is efficient for improvement of SNP data classification, especially the new data selected and generated by CBFS + FFM and classified by SVM tend to produce the most favorable accuracies.

Acknowledgments. This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2012S1A2A1A01028576).

References

1. Waddel, M., Page, D., Zhan, F., et al.: Predicting cancer susceptibility from single-nucleotide polymorphism data: a case study in multiple myeloma. Life and Medical Sciences (2005)
2. Kira, K., Rendell, L.A.: The feature selection problem: traditional methods and new algorithm. In: Proceedings of AAAI (1992)
3. Dutoit, S., Fridly, J.: Introduction to classification in microarray experiments. A practical approach to microarray data analysis, pp. 132–149 (2003)
4. Dy, J.G.: Unsupervised feature selection. Computational methods of feature selection, pp. 19–39 (2008)
5. Liang, J., Yang, S., Winstanley, A.: Invariant optimal feature selection: A distance discriminant and feature ranking based solution. Pattern Recognition 41, 1429–1439 (2008)
6. Robnik-Sikonja, M., Kononenko, I.: Theoretical and empirical analysis of ReliefF and RReliefF. Machine Learning 53, 23–69 (2003)
7. Lee, J., Batnyam, N., Oh, S.: RFS: Efficient feature selection method based on R-value. Computers in Biology and Medicine (2012)
8. Seo, M., Oh, S.: CBFS: High performance feature selection algorithm based on feature clearness. PLoS ONE 7(7) (2012)
9. Cover, T., Hart, P.: Nearest Neighbor pattern classification. IEEE 13(1), 21–27 (1967)
10. Chang, C., Lin, C.: LIBSVM – A library for support vector machines (2005), `http://www.csie.ntu.edu.tw/cjlin/libsvm/`
11. Seo, M., Oh, S.: Derivation of an artificial gene to improve classification accuracy upon gene selection. Computational Biology and Chemistry 36, 1–12 (2011)

12. Barret, T., Edgar, R.: Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. Methods in Enzymology, 352–369 (2006), http://www.ncbi.nlm.nih.gov/geo/
13. Marshall, C.R., et al.: Structural variation of chromosomes in autism spectrum disorder. Am. J. Hum. Genet. 82(2), 477–488 (2008)
14. McMullan, D.J., et al.: Molecular karyotyping of patients with unexplained mental retardation by SNP arrays: a multicenter study. Hum. Mutat. 30(7), 1082–1092 (2009)
15. Reid, J.F., et al.: Integrative approach for prioritizing cancer genes in sporadic colon cancer. Genes Chromosomes Cancer 48(11), 953–962 (2009)
16. Katoda, M., et al.: Identification of novel gene amplifications in breast cancer and coexistence of gene amplification with an activating mutation of PIK3CA. Cancer Research 69(18), 7357–7365 (2009)
17. Evans, D.T.: A SNP microarray analysis pipeline using machine learning techniques. M.S., Computer Science, Ohio University (2010)
18. Oh, S.: A new dataset evaluation method based on category overlap. Computers in Biology and Medicine 41, 115–122 (2011)
19. Mukherjee, S.: Classifying microarray data using support vector machines. A practical approach to microarray data analysis, pp. 166–185 (2003)
20. Batnyam, N., Tay, B., Oh, S.: Boosting classification accuracy using feature fusion. In: 2012 International Conference on Information and Network Technology (ICINT), vol. 37 (2012)
21. Hanczar, B., Zucker, J.D., et al.: Feature construction from synergetic pairs to improve microarray-based classification. Bioinformatics 23, 2866–2872 (2007)