Disease Single Nucleotide Polymorphism Selection using Hybrid Feature Selection Technique

Manu Phogat1*, Dr. Dharmender Kumar1
1Guru Jambheshwar University of Science and Technology, Hisar, Haryana, India.

*Email: Kunjean4181@gmail.com

Abstract. According to recent studies the Single Nucleotide Polymorphism (SNPs) plays very important role as genetic marker in various complex diseases. Lots of machine learning techniques are already applied on SNPs data to distinguish between affected and healthy individuals. The major problem with the SNPs dataset is high number of features and small number of samples which are referred as ‘large p’ and ‘small s’ problem. In this paper we proposed a hybrid feature selection method for selecting an optimal subset of SNPs and from that we select the significant SNPs, which act as marker for disease. The method is a hybrid technique based on combination of filter and wrapper method, the (mRMR) Minimum Redundancy Maximum Relevancy and Particle Swarm Optimization for Gene Selection with Support Vector machine (PGOGS-SVM) respectively. The proposed mRMR+PSOGS-SVM approach has been applied to mental retardation SNP dataset taken from NCBI-GEO website. The method has achieved high classification accuracy up to 88% and outperformed all other compared feature selection techniques.

Keywords: Feature selection, Mental Retardation (MR) Machine learning, Single Nucleotide Polymorphism (SNP), Support Vector Machine, Hybrid Techniques.

1. Introduction
Approximately there are 3 billion base pairs in human genome called as nucleotides. The difference in single nucleotide in the DNA is referred as Single Nucleotide Polymorphism, which is the cause of most common genetic variation amongst the people. The SNPs generally occurs throughout the human genome and they are estimated to be present once in every 1000 base pairs. The SNPs are genetic variants between individuals that are used as an efficient marker (genes) for many complex diseases. It has been reported that the SNPs cause the phenotypic changes like height, colour, skins etc. and the simplest example is sickle cell anaemia [1]. Many complex diseases are having combination of environment factor and genetic factors. The Gene Wide Association Study (GWAS) has been recently used to understand the relationship between genetic variation and diseases [2].

The current value predictions of SNPs are 11 million [3], as per the highly abundant microarray platform, the genotype has approx. 1 million SNPs [4]. To predict SNPs interaction within complex diseases is not an easy task, so to overcome with this problem, Machine Learning techniques are commonly applied on SNPs data sets for classification. The main challenge for SNPs dataset is that there is very small number of samples (few hundred) but large number of SNPs (about to millions), this is referred as the “The Curse of Dimensionality” problem [5]. The main target of SNPs analysis is to build models that distinguish the healthy individual from the affected one with the disease. So, the selection of SNPs subset that is highly discriminative is crucial and this problem is called feature selection [6]. The main objective of feature selection technique is to minimize the feature subset and maximize the classification accuracy. There are multiple dimensionality reduction techniques that
have been usually applied to transform the initial feature set into a smallest subset associated with disease [7]. Several revolutionary methods have been already used in different gene studies [8][9]. For the success of a Machine Learning model the selection of good FS method is very crucial. By definition the process of removing irrelevant features and create a smaller feature subset so that the machine learning technique’s select only those features of training data that are useful for production and analysis is called feature selection [10][11]. The analysis of SNPs data sets is playing a crucial role in disease – gene association studies. Various feature selection techniques have been used to select the most informative SNPs. Batanyam et.al [12] used a three-stage approach for the classification of the SNP data, first they applied various feature selection techniques like RFS, FSDD, ReliefF to select most informative SNPs- then they used (FFM) feature fusion method for generating new artificial features of selected SNPs, finally they did the classification task. In various studies usually there are three types of feature selection methods which have been used: Filter, wrapper and Embedded method [13]. The filter methods are learner’s independent method. They rely on characteristics of data. The filter methods select the features based on different criteria such as mutual information, symmetrical uncertainty and fisher score. The main advantage of filter method is that they are fast, efficient and more scalable. As they are not interacting with classifier, so they usually have low performance results. The wrapper approach basically depends on the learner (Classifier) to know the effect of various feature subsets [14] [15]. The wrapper method searches for a subset of feature and evaluate the selected feature via learner’s performance. They offer better classification accuracy than the filter methods. The main disadvantage is that they are computationally expensive for large data sets like SNPs. So, there is an intermediate solution called as the embedded methods to overcome the problem of both filter and wrapper methods. The embedded methods include the advantages of both methods, recently many embedded methods have been used usually like first filter method applied and then followed by the wrapper method [16].

Anekboon et al. [17] proposed a hybrid method such as CBFS+KNN and ANN+RR, where first filter approach is used then wrapper approach. The researchers avoid using wrapper and filter technique alone in SNPs analysis because of higher dimensionality of the data and less accuracy. In the present paper we have proposed a hybrid method using mRMR as a filter approach to select the subset of SNPs, followed by PSO-SVM as wrapper -feature selection technique. The proposed method has been applied on Mental Retardation dataset retrieved from NCBI (GEO). The paper has been divided in three sections. Section 2 describes the proposed feature selection method and classification task. However, section 3 discusses the experimental results followed by conclusion in section 4.

2. The Proposed structure
The proposed structure is for analysing mutation related complex disease as described in figure1. This includes the three-stage framework: the pre-processing followed by the transformation of data and to find out the most relevant feature that are associated with disease by propose feature selection technique and the last step is classification of the relevant feature.

![Figure 1. Data flow diagram for proposed technique.](image-url)
2.1 Pre-Processing
The pre-processing stage is conducted into two sub stages.

1. Transformation of the SNPs data.
2. Data refinement.

The SNPs data are usually in character string so this makes difficult for analysis process. The data needs to be converted into numerical form. This process can be applied in many ways [18] [19]. In the first step of pre-processing, we have used the transformation process used by Naeem et.al [20] which is a lossless transformation and the SNPs values are converted from {AA, AB, BB, NC} to {11,01,10, and 00} as shown in figure 2.

![Table 1](image1.png)

**Figure 2.** Transformation of SNPs datasets, where s is the number of features and x the number of samples

2.2 Data Refinement

**Removals of duplicate SNPs**- The duplicate SNPs are those, which have same values in case and control samples, for example a SNP having BB value in both cases and control. The MR Dataset had total 262328 SNPs, out of which, 1% duplicate SNPs were removed that reduced the total number of SNPs to 260317. Removing of such redundant SNPs reduce the data size and maintain efficient storage.

**No Call Replacements**- the No Call values of SNPs data are produced when the value of allele is not finding by the DNA sequencer, the No call value is marked as “NC”. The SNPs which have more than 10% of “NC” values are discarded.

The other remaining “NC” values in the SNPs are replaced by most relevant value. Okser et.al and Wuet.al [21] [22] have widely used this technique. MR data set contains large number of NC values, more than 12% of total SNPs are having more than or equal to 10% “NC” value, after pre-processing the remaining SNPs in MR data set reduced to 221507.

2.3 Proposed Feature Selection Technique
After pre-processing of data, a feature selection technique is needed to select the most relevant subset of features. In this paper mRMR and PSOGS-SVM algorithm have been fused. Where mRMR is used as a filter approach and PSOGS-SVM is used as a wrapper approach. Brief explanations about these approaches are given below.
2.3.1 Minimum Redundancy Maximum Relevance (mRMR)

The mRMR technique was introduced by Ding and Pang [23]. It is a very effective and fast filter feature selection technique for genomic data. The mRMR algorithm selects important features by minimum redundancy and maximum relevance based on mutual information between features. The minimum redundancy means to select the genes that are mutually dissimilar. The minimum redundancy is expressed in equation 1.

$$\min_{W_i} W_I = \frac{1}{|S|^2} \sum_{l,j \in S} I(l, j),$$

where S is set of features, I(i,j) is mutual information between feature i& j.

The maximum relevance condition is used mutual information between gene and target class to maximize the total relevance of all genes. The maximum relevance condition is expressed as.

$$\max_{V_i} V_I = \frac{1}{|F|} \sum_{i \in F} I(h, i)$$

The first feature in mRMR is selected according to the equation 2, first the feature with the highest I(h,i) is select and rest are selected incremental way.

2.3.2 PSOGS-SVM

PSO was introduced in 1995 by Kennedy and Eberhart [24]. It is based on the social behaviour of flocking of bird or fish schooling. The PSO-SVM is a filter-based method given by Xue et.al [25]. In this article we have proposed a Particle Swarm Optimization for gene selection (PSOGS) method as a wrapper approach to select the efficient SNPs for classification of diseases. The basic functionality of PSOGS algorithm is as follows:

1. Initialization of the particles which are randomly generated. To find the optimal solutions all particles move around in problem space.
2. Each particle in the population should be measured on the basis of its fitness.
3. Evaluate the velocity function of each particle at the t-th iterations according to following equations:

$$V_i(t + 1) = w \times V_i(t) + c_1 r_1 \times (pbest_i(t) - x_i(t)) + c_2 r_2 \times (gbest(t) - x_i(t))$$

Where position vector $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, velocity vector $V_i = (v_{i1}, v_{i2}, \ldots, v_{in})$, $x_{id}$ is a binary bit, $i=1, 2, \ldots, m$, $d=1, 2, \ldots, n$, $m$ = total no of particles, $n$ = dimension of the data. $C1$ and $C2$ are acceleration constants with the value between [0, 2], $w$ is the inertia weight and $r1$, $r2$ and $r3$ are the random numbers value between [0,1].

4. Each particle moves to the next position.

$$S \text{ig}(V_i(t + 1)) = \frac{1}{1 + e^{-v_{i}(t+1)}}$$
If \( \text{Sig}(V_i(t+1)) > r3 \) then \( x^d_i(t+1) = 1 \); (gene selected)

Else \( x^d_i(t+1) = 0 \) (gene discarded)

The sigmoid function \( \text{Sig} \left( V_i^d(t+1) \right) \in [0,1] \) (7)

To find out the importance of a particle (gene subset) we calculate the fitness value.

\[
\text{fitness}(X_i) = w_1 \ast A(X_i) + \left( \frac{w_2(M - R(X_i))}{M} \right)
\]

(8)

\( M \) = total no of genes for each sample, \( R(X_i) \) = No. of gene selected in \( X_i \). \( w_1 \) and \( w_2 \) are priority weights where \( w_1 \in [0.1,0.9] \) and \( w_2 = 1 - w_1 \). While \( A(X_i) \in [0,1] \) is LOOCV (Leave One out Cross Validation) accuracy that use only the genes in gene subset(\( X_i \)). The accuracy is defined by SVM classifier.

5. If termination criterion meets, then stop the algorithm otherwise return to step 2.

In PSOGS feature selection, if number of features are chosen are \( n \) then there has to be \( 2 + n \) decision variables and the value of \( n \) ranges between 0 and 1. If the value of the bit vector is 1 then the gene is selected and if its value is 0 that means the gene is discarded.

Table 1: The PSOGS parameters.

| Parameters               | value |
|--------------------------|-------|
| Swarm size               | 50    |
| Inertia Weight           | 0.9   |
| Acceleration constants C1, C2 | 2     |
| Maximum No. of iterations| 30    |

The SNPs dataset have huge number of features, which makes the feature selection very necessary to reduce the feature space and improve the accuracy. Here we have designed a hybrid feature method by combining mRMR and PSOGS+SVM techniques. Our FS method overcomes the disadvantages of mRMR and other wrapper-based techniques. The mRMR is a filter-based technique so the classifiers are not included in the first step. mRMR has limited performance when we use it alone, and similarly PSOGS+SVM is not applicable on enormous data sets such as SNPs, which are having a large feature space in millions. The PSOGS+SVM are very time consuming when we apply it alone. So, we enhance the selection process in PSOGS+SVM by using mRMR filter approach at first, so this technique improves the classification accuracy. The workflow of algorithm is presented in figure 3. After pre-processing SNPs data set, the filter method mRMR is applied. The mRMR work on mutual information between the target class and the attributes variables. The \( n \) number of feature subset, which is selected by mRMR techniques gives as an input to wrapper based PSOGS+SVM algorithm. The first step in PSOGS+SVM is the initialization of the population of particle where each particle has been assigned a random position within \( D \) dimension space having a relative velocity for each particle. In the next step the fitness of each particle is evaluated for SVM. For each particle, fitness is considered to be classification accuracy in the study. If the fitness is better than the best fitness of the particle, then the positive vector is saved for the particle. If the fitness is better than the global best than the global best position vector is saved. The position and velocity of particle are updated until the termination criterion is satisfied.
2.4 SNPs Datasets
The proposed work used the Mental Retardation SNP microarray dataset which is available publicly at NCBI with GEO accession GSE13117 and platform GPL3718. The NCBI has free public repository named as GEO (Gene Expression Omnibus) that has numerous amounts of datasets like NGS, SNPs and other genomic microarray data of various diseases. In mostly datasets the samples are labelled in two classes such as case and control, the control samples are healthy ones and case samples are affected ones. All samples both case and control are having a unique identification number and a sequence of SNPs alleles. The SNP marker is of two types homozygous and heterozygous, the values AA and BB is for homozygous allele and AB for heterozygous allele. The MR Dataset consists of 262328 SNPs (features) and 360 samples, in which 120 are case and 240 are control samples [26].
3. Results and Discussion
The proposed mRMR+ PSOGS-SVM was implemented in MATLAB package version r 2018a. The classifier ANN, SVM, NB and LDA are applied with LOO-CV for obtaining results. The two measures used for prediction the rate of classification are accuracy (Acc) and F-measure (F).

\[
Acc = \frac{TP+TN}{TP+FP+FN+T} \times 100
\]

\[
F = 2 \times \frac{Pre \times Re}{Pre + Re}
\]

When

\[
Re = \frac{TP}{TP+FN} \times 100
\]

\[
Pre = \frac{TP}{TP+FP} \times 100
\]

TP = True Positive, TN= True negative, FP= False Positive, FN= False Negative

Table 2: The performance of different classifier on MR data set

| MR          | SVM | ANN | NB | LDA |
|-------------|-----|-----|----|-----|
| mRMR        | 81.67 | 74.72 | 80.00   | 76.67   |
| CMIM        | 82.83 | 69.44 | 69.76   | 32.44   |
| ReliefF     | 76.72 | 69.44 | 81.93   | 73.92   |
| CMIM RFE    | 81.88 | 66.31 | 82.99   | 70.45   |
| mRMR_PSOGS  | 81.88 | 66.11 | 65.28   | 63.16   |

An optimal subset of 50 SNPs was used for obtaining the experiment results. Table 1 shows the performance measures used are accuracy and F-measure to compare proposed method with ReliefF, CMIM, mRMR and CMIM – SVMRFE and it clearly shows that the proposed FS method outperforms the compare methods on given data set.

By comparing the obtained result our method shows superiority over compared algorithms in the tested data sets with 5% increase in average accuracy. The result shows that mRMR-PSOGS+SVM hybrid method gives good performance over all classifier except KNN that achieved low accuracy out of all the classifiers. The figure 4 shows that the proposed algorithm is validate and about to reach best accuracy with feature subset is more than 20 SNPs.
Figure 4. Performance comparison between proposed and other methods over (10, 20, 30, 50, 70 and 100) SNPs.

The best accuracy was obtained while using 50 SNPs and the method outperforms CMIM, Relief F, mRMR and CMIM – SVMREF methods. The observation shows that the proposed method obtained up to 87% accuracy with 50 SNPs, which is very useful for medical diagnosis.

Table 3: The comparison of classification accuracy of different FS method including proposed method on MR data sets.

| MR       | SNPs | Acc% | Ref |
|----------|------|------|-----|
| DS+SVM   | 6    | 59   | [2] |
| Chi+SVM  | 2    | 59   | [2] |
| CMIM+SVM | 100  | 83   | [12]|
| mRMR+SVM | 100  | 82   | [12]|
| ReliefF+SVM | 30  | 78   | [12]|
| CMIM+SVMREF | 50  | 85   | [1] |
| Proposed (SVM) | 50 | 88   |     |

The chi+SVM and DS+SVM are having worst performance with approx. 59% accuracy using only 2 and 6 SNPs. The present method achieved highest accuracy among all the FS methods with a subset of 50 SNPs.

To examine the statistical significance of the results a non-parametric Kruskal-Wallis H test is used. The result comparison of the proposed method and other methods show in table with P value (P<0.05).
Table 4. KRUSHKAL WALLIS H Test for Statistical significance between proposed and other methods.

| SVM | KNN | LDA | NB |
|-----|-----|-----|-----|
| RF  | mRMR | CM, RFE | Pro | RF  | mRMR | CM, RFE | Pro | RF  | mRMR | CM, RFE | Pro |
| SVM+Proposed | + | + | + | 1 | + | + | + | 1 | + | + | + | 1 | (+) states a significance with p ≤ 0.05 and (-) states the significance p > 0.05. |
| KNN+Proposed | + | - | + | + | + | 1 | + | - | + | + | 1 | + | + | + | + | |
| LDA+Proposed | + | + | - | + | + | + | 1 | + | + | - | + | + | + | + | |
| NB+Proposed | + | + | + | + | + | - | + | + | + | + | - | + | + | + | - | |

As shown in the above-mentioned results the proposed algorithm outperforms the all compared algorithms in terms of accuracy and statistical significance.

4. Conclusion

The proposed work presented a hybrid approach using mRMR+PSOGS-SVM model to deal with the high dimensional datasets especially the genomic data like SNPs datasets. The technique is a two-stage approach, first the mRMR filter technique used to select the best subset form the large feature space. Then we applied the PSOGS-SVM method that not only discarded the redundant genes but also improved the accuracy of the model which helps in predicting the complex disease. The experiment conduct on MR Dataset obtained from NCBI (GEO) database and the proposed technique outperformed the compared methods (CMIM, mRMR, ReliefF, CMIM-SVMRFE). Using different number of SNPs on tested datasets up to 88% accuracy was achieved. From the derived results we conclude that, to distinguish the affected person with complex diseases the SNPs dataset of whole genome can be used.

References

[1] M. J. Stuart and R. L. Nagel, 2004 “Sickle-cell disease,” Lancet, vol. 364, no. 9442, pp. 1343–1360.
[2] D. T. Evans, “A SNP Microarray Analysis Pipeline Using Machine Learning Techniques,” Ohio University, 2010.
[3] K. A. Frazer, S. S. Murray, N. J. Schork, and E. J. Topol, 2009 “Human genetic variation and its contribution to complex traits,” Nat. Rev. Genet., vol. 10, no. 4, pp. 241–251.
[4] P. M. Visscher and G. W. Montgomery, 2009 “Genome-wide Association Studies and Human Disease,” JAMA, vol. 302, no. 18, p. 2028.
[5] M. Waddell, D. Page, and J. Shaughnessy, 2005 “Predicting cancer susceptibility from single-nucleotide polymorphism data,” in Proceedings of the 5th international workshop on Bioinformatics - BIOKDD ’05, p. 21.
[6] M. Y. MohdSaberi Mohamad, SigeruOmatu, 2009 “Cyclic hybrid method to select a smaller subset of informative genes for cancer Classification,” Int. J. Innov. Comput. Inf. Control, vol. 5, no. 8, pp. 2189–2202.
[7] H. Schwender and K. Ickstadt, 2008 “Identification of SNP interactions using logic regression,” Biostatistics, vol. 9, no. 1, pp. 187–198.
[8] J. M. Deutsch, 2003 “Evolutionary algorithms for finding optimal gene sets in microarray prediction,” Bioinformatics, vol. 19, no. 1, pp. 45–52.
[9] T. Jirapech-Umpai and S. Aitken, 2005 “Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes,” BMC Bioinformatics, vol. 6, no. 148, pp. 1–11.
[10] N. Batnyam, A. Gantulga, and S. Oh, 2013 “An Efficient Classification for Single Nucleotide Polymorphism (SNP) Dataset,” in Studies in Computational Intelligence, Springer International Publishing Switzerland, pp. 171–185.
[11] N. Batnyam, A. Gantulga, and S. Oh, 2013 “An Efficient Classification for Single Nucleotide Polymorphism (SNP) Dataset,” in Studies in Computational Intelligence, Springer International Publishing Switzerland, pp. 171–185.
[12] M. Seo and S. Oh, 2012 “CBFS: High Performance Feature Selection Algorithm Based on Feature Clarity,” PLoS One, vol. 7, no. 7, p. e40419.
[13] I. H. Witten, E. Frank, and M. a Hall, 2011 Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed. Morgan Kaufmann.
[14] R. Ruiz, J. C. Riquelme, and J. S. Aguilar-Ruíz, 2006 “Incremental wrapper-based gene selection from microarray data for cancer classification,” Pattern Recognit., vol. 39, no. 12, pp. 2383–2392.
[15] S. Uppu, A. Krishna, and R. P. Gopalan, 2018 “A Review on Methods for Detecting SNP Interactions in High-
Dimensional Genomic Data,” *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 15, no. 2, pp. 599–612.

[17] K. Anekboon, C. Lursinsap, S. Phimoltares, S. Fucharoen, and S. Tongsima, 2014 “Extracting predictive SNPs in Crohn’s disease using a vacillating genetic algorithm and a neural classifier in case-control association studies,” *Comput. Biol. Med.*, vol. 44, no. 1, pp. 57–65.

[18] V. Lagani et al., 2009 “A novel similarity-measure for the analysis of genetic data in complex phenotypes,” *BMC Bioinformatics*, vol. 10, no. Suppl 6, p. S24.

[19] H.-J. Ban, J. Y. Heo, K.-S. Oh, and K.-J. Park, 2010 “Identification of Type 2 Diabetes-associated combination of SNPs using Support Vector Machine,” *BMC Genet.*, vol. 11, no. 1, p. 26.

[20] R. Alzubi, N. Ramzan, H. Alzoubi, and A. Amira, 2018 “A Hybrid Feature Selection Method for Complex Diseases SNPs,” *IEEE Access*, vol. 6, pp. 1292–1301.

[21] T. Pahikkala, S. Oksa, A. Airola, T. Salakoski, and T. Aittokallio, 2012 “Wrapper-based selection of genetic features in genome-wide association studies through fast matrix operations,” *Algorithms Mol. Biol.*, vol. 7, no. 1, p. 11.

[22] D. Yao, J. Yang, X. Zhan, X. Zhan, and Z. Xic, 2015 “A novel random forests-based feature selection method for microarray expression data analysis,” *Int. J. Data Min. Bioinform.*, vol. 13, no. 1, p. 84.

[23] C. Ding and H. Peng, 2005 “Minimum redundancy feature selection from microarray gene expression data,” *J. Bioinform. Comput. Biol.*, vol. 03, no. 02, pp. 185–205.

[24] J. Kennedy and R. Eberhart, “Particle swarm optimization,” in *Proceedings of ICNN’95 - International Conference on Neural Networks*, 1995, vol. 4, pp. 1942–1948.

[25] B. Xue, L. Cervante, L. Shang, W. N. Browne, and M. Zhang, 2012 “A multi-objective particle swarm optimisation for filter-based feature selection in classification problems,” *Conn. Sci.*, vol. 24, no. 2–3, pp. 91–116.

[26] D. J. McMullan et al., 2009 “Molecular karyotyping of patients with unexplained mental retardation by SNP arrays: A multicentre study,” *Hum. Mutat.*, vol. 30, no. 7, pp. 1082–1092.