Microarray based Multi Filter Fusion Gene Selection and Ensemble Classification of Leukemia Sub Types

C.L.Edirimanna[1], T.M.K.K. Jinasena[2], and E.A.T.A.Edirisuriya [3]
[1] *Faculty of Information Technology, University of Moratuwa,Sri Lanka*
[2,3]*Department of Computer Science,University of Sri Jayawardanapura, Sri Lanka*
[#]*lakminiedirimanna@gmail.com*

Abstract— *Leukemia is a blood cancer which exists in bone marrow. Two major acute leukemia types, Acute Lymphocytic Leukemia (ALL) and Acute Myelogenous Leukemia (AML) need immediate treatments .Conventional lab methods take more time to differentiate these two types risking the patient's life. The invention of the micro array technology has been recognized as a major advancement in cancer diagnosis and prognosis. However, these gene expression data has a significant higher number of dimensions. This curse of dimensionality makes it difficult to find associations and patterns across multiple dimensions. The benchmark micro array data set, consisting of 72 patients with 7000 attributes, has been used. Extracting genes only related to the disease and classifying them across the multiple dimensions are the research challenges. A Multi Filter Fusion based gene selection and an Ensemble based Classifier (MFF-EC) is proposed to improve the accuracy of individual filters. The main three steps are (1) Feature selection (2) Multi filter fusion and (3) Ensemble classification of ALL/AML. Further, both parallel and sequential approaches are used for step 1 and 2 separately. ReliefF, Correlation, Gain Ratio, and Weight-Support Vector Machine are used as distinct feature selection methods. In MFF-EC parallel approach, a consensus score is introduced to select a sub set of genes from each individual filter method and combined the maximum scored genes. In MFF-EC sequential approach, selected filters are applied one after the other to gradually reduce the number of dimensions. Finally, an ensemble classifier is used to combine the results of multi filters. Performances of the classification models have been evaluated and MFF-EC parallel performs better than the other four methods in terms of accuracy, sensitivity and specificity with average values 98.56, 98.87 and 99.1 respectively.*

Keywords— **BioInformatics, Leukemia, Multi Filter**

I. INTRODUCTION

Cancer has become one of the major causes of morbidity and mortality worldwide. World Health Organisation (WHO) has reported that, in 2012, approximately 14 million new patients and 8.2 million cancer related deaths have been recorded and it is expected that the number of newly identified patients will rise by about 70% within the next two decades (WHO). Therefore, early detection of cancer is a crucial, not only for increasing the survival rate, but also for cancer diagnosis, prognosis, therapeutic and drug discovery process.

Leukemia is a blood cancer which exists in bone marrow. Many researchers have reported that leukemia have several sub types which vary from to each other (Ghosh et al.2012). Thus classification of different sub types of leukemia is greatly important in leukemia diagnosis and prognosis. Morphological appearances of tumours are the basic feature to cancer classification (Etzioni et. Al. 2001). However, this usual technique is unable to differentiate tumours with similar histopathology features (Singer et.al. 2001).So the limitations of the conventional methods led to tumour classification from morphologic methods to molecular methods. The invention of the micro-array technology has been recognized as major advancement in classification of cancers. Thus micro-array technology has the ability of screening a large number of gene expression profiles at once (Singer et.al. 2001).

Filter technique is the earliest method that evaluates the gene based on intrinsic characteristics of the gene. Filter method is characterized with better generalization property since gene selection is independent of any classifier method (Yukyee et.al.2010). Most common method of selection of informative genes list is that genes are obtained a score according to the feature selection algorithm. Then the genes with the highest scores are selected as informative genes list. Main disadvantage of this individual filter method is that they may leave out some informative genes. In literature, hybrid approaches have been overcome by these drawbacks of individual filter methods with acceptable performance (Huereta et.al. 2010).However these hybrid approaches has shown that the performance of the classifier depends on the choice of the feature selection methods. Consensus of filter methods seems to be a solution that overcomes this limitation of single filter methods.

Consensus of multi filter criteria means that the selection of a sub set of genes that represent aggregation of result of filter methods into one rationale. In order to improve the accuracy of classification, a method has been proposed in this paper by combining strength of multi filter criteria consensus and ensemble technique.

Rest of the paper is organized as follows. Section 2 describes a brief summary of the existing multi filter approaches. Section3 gives the design of the proposed system and evaluation method describe in section4. Finally section5 presents the conclusion and further works.

II. EXSISTING MULTI FILTER APPROACHES

This section summarizes the existing multi filter approaches and how they combine the individual filters. It further analyses methods of the results obtained through the multi filters.

Table 1.Summary of the few existing multi filter approaches and proposed method

Reference	Used individual filter methods	Filter combination method	Further analysis
Yukee et.al. 2010	SNR , t-statistic , Pearson's correlation combination	Union of the top ranked features of each individuals	Wrapper method
Mao et.al. 2011	Fisher's Ratio , Relief , SVM, ADC	Score combination followed by rank combination	SVM –REF
Huerta et al. 2015	MI , SNR , Wilcoxon test , BSS/WSS , t-statistics	Rank aggregation followed by score calculation and ranking	Embedded
Shreem et al. 2013	Relief , mRMR	Sequential filter aggregation	Genetic Algorithm
Rathore et al. 2014	Chi-square , F-Score , PCA , mRMR	Do not aggregation, only validation	Ensemble classificatio n (simple voting)
Alonso-Betanzos et.al 2014	CFS , Consistency filter, INTERACT, Information Gain , ReliefF	E1 approach - don't combine filters- E2 -select sub set of feature which outperform the classifier	-E1 – ensemble classifiers- -E2-classify with basic two classifiers
Dittman et.al 2013	25 filter methods in 3 category (Threshold based , Statistical based, commonly used)	-Mean -Median -Enhanced borda -Exponential weigh -Highest rank -Lowest rank -Robust Rank -Round robin -Stability selection	Classify with five basic classifiers and comparison
Sasikala et.al. 2014	PCA (Feature Extract), CFS , Symmetrical - Uncertainty	Sequential filter aggregation	Classify with four basic filters
Propose method	Releif , CFS , Gain Ratio, SVM	-Parallel – Score based -Sequential – Multi filtration	Ensemble classificatio n (AdaBoostM 1 and Bagging with NB and Ibk basic classifier)

III. METHODOLOGY

Multi filter fusion based gene selection and ensemble based classifier (MFF-EC) algorithm has been designed in three unique stages with two approaches; Multi Filter Fusion parallel (MFF-parallel) and Multi Filter Fusion sequential (MFF-sequential).

Main three stages are as follow;

(1) Feature selection using four distinct filter methods

(2) Multi filter fusion

(3) Ensemble classification of ALL/AML

In proposed algorithm, last stage is common for both approaches while the other two stages differ on how filters integrate. Fig1 represents the overall design diagram of the MFF-EC.

A. Feature Selection using Four Distinct Filter Methods

The basic distinct filter criterions used in this study are ReliefF, Correlation-based Feature Selection (CFS), Gain Ratio and Absolute Weight–Support Vector Machine (AW-SVM). ReliefF is a multivariate, distance based feature selection measure while CFS is multivariate, and correlation based feature selection measure. Gain Ratio is a univariate, probabilistic based feature selection method while AW-SVM is embedded method that concerns on features coefficient in SVM. The key idea

behind the use of these filters is to obtain informative gene set by each filter. In parallel approach, each distinct filter selects a set of genes and combines them across a feature scoring method. Proposed sequential method also used the same filters and sequentially combined over with best filter criteria arrangement.

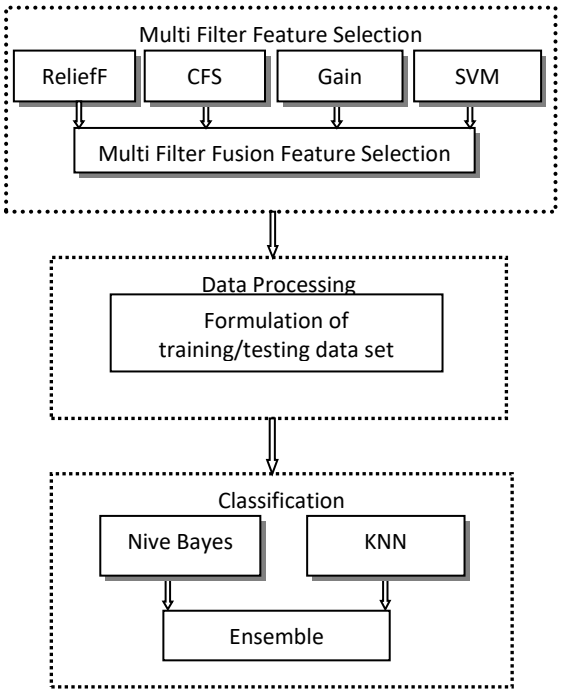

Fig 1. Semantic representation of MFF -EC

B. Multi Filter Fusion

1) Parallel Approach: This approach in MFF-EC method used four different sub set of gene list which is selected at stage one. Above filter techniques are representing different feature evaluation criteria in filter paradigm. Most of them are assigned a rank for each gene base on the score which is calculated according to the different gene evaluation criteria.

Gene score calculation is a main component in parallel approach. Selected individual filter techniques assign a score for all genes based on some criteria which vary from one method to another. Thus, MFF-EC proposes a consensus score based method to select new informative gene set by aggregating gene score which has been assigned by individual filter criteria.

Assume f_{score} is the filter score obtained by a gene from basis filter criteria, n is the number of basis criteria used in fusion and w_i is weight assigned by each filter method. In this research we assume that $w_i = 1$ and consensus gene score is defined as follow:

$$Gene\ Score = \frac{1}{n}\sum_{i=1}^{n} w_i\ f_{score} \quad (1)$$

Then, genes which obtained highest score as gene score forms a new informative gene list which is used to improve the classification accuracy. However in some cases, algorithms are disabling to access on filter score while displaying gene rank as the results. In such scenarios, MFF-EC proposes a common rank based method to select informative genes from pre processed genes which has been ranked based on individual filter criteria. Assume g(r) is the rank list produced by individual filter criteria r; all genes of feature set filtered by criteria r assigns a score as follows:

$$Score = 1 - \left(\frac{g_r - g_{r\,min}}{g_{r\,max} - g_{r\,min}}\right) \quad (2)$$

where $g_{r\,max}$ and $g_{r\,min}$ are the maximum and minimum value in genes list. Thus maximum score obtained by a particular gene is 1 for one particular criterion while others are acquires score less than one relevant to their rank in a sub set of gene.

Then gene score is calculated by simple but effective way of score combination method by taking the average of the score which obtain by equation 2. For a particular gene g filtered by individual filter criteria r, MFF-parallel gene score is calculated as follows.

$$Gene\ score = \frac{1}{m}\sum_{i=1}^{m} g_{score} \quad (3)$$

where m is the number of filter criteria while g_{score} is obtained from equation 2.

MFF-parallel approach has been implemented as a separate feature selection method using java based WEKA API.

2) Sequential Approach: Sequential approach in the MFF method combines four selected different filter methods over best combination of the filter criteria in sequential manner. Genes are filtered by one filter by other. The boundary between one filter to another is decided on parameters called threshold value, best set of selected genes from one filter method or natural boundary indication. Then several experiments have to be conducted to obtain a best sequential combination of the filters over above parameters. The best model is designated with best classification accuracy with tested classifier. With the time limitation, we have only used two individual filter methods to arrange in sequential manner. Table2 shows that used filter arrangements.

C. Ensemble Classification of ALL/AML

A classifier ensemble is an integration of classification models, referred to as base classifiers. Individual decisions of classifiers are combined in order to obtain a final prediction. The aim of using a classifier ensemble is to provide an overall level of performance which is superior to the performance of any of the single base classifiers. The advantages of ensemble classification such as less prone to over fitting and improve the classification performance (Dittman et al.2013). In this research, Naïve bayes and K-NN classifiers employ as basic classifier connected with ensembles in sequential manner to improve the classification accuracy. AdaBoost and Bagging is the ensemble techniques which are used to design the powerful classifier (Abeel at.el.2010). A classification model has been proposed with MFF feature selection and ensemble of naive bayes and K-NN classifiers.

Table 2: Sequential filter arrangements

	Order of the sequential filter arrangement	
Case 1:	SVM	Gain
Case 2:	SVM	CFS
Case 3:	SVM	Relief
Case 4:	Relief	SVM
Case 5:	Gain	SVM
Case 6:	CFS	SVM

IV. RESULTS AND EVALUATION

Accuracy of the classifier is evaluated using various criteria. Benchmark leukemia data set was obtained from broadinstitute.org data repository which has a huge number of various data sets (Alonso-Betanzos at el .2014). The selected data set contains 72 samples and 7129 attribute with two distinct values for class attributes. The 72 of micro array dataset consists of 25 samples of AML, and 47 samples of ALL. Performance of the classification models are evaluated in terms of the sensitivity, specificity, test classification error and accuracy.

D. Performance Evaluation of MFF-EC on Classification Error Rate

According to the test classification error results in Table 3, MFF sequential and parallel approaches have obtained the lowest error when compared to individual filter methods excluding CFS and SVM. The best test classification error among the basic classifiers was represented by one MFF sequential approach of SVM-CFS combination with NB classifier. With the value of 0.12, MFF- parallel approach obtained best result among the ensemble classifiers (AdaBoostM1-NB).

When compared to MFF-parallel with basic classifier values to ensemble classifier values, Ibk and NB both have improvement with ensembles. Then MFF-sequential approach obtains better results with basic classifier when compared to ensemble of classifier. Thus among the four ensemble classifiers (AdaBoostM1-NB, AdaBoostM1-IBk, Bagging-NB and Bagging-Ibk) MFF-parallel obtained best results for three ensembles and other ones win by MFF-sequential (SV-CF).

E. Performance Comparison of MFF-EC with Existing Multi Filter Schemes and Classifier

Table 4 represent the comparison of test classification results between filter ensembles methods in existing reference (Alonso-Betanzos at el .2014) and proposed MFF-EC approaches. Proposed method and existing method both have been used Ibk and NB classifier. For both classifiers, existing method has been shown that same test classification error. As shown in Fig2, MFF – parallel obtains the best results for Ibk while MFF – sequential (SVM-CFS) obtains best and overall minimum result for NB.

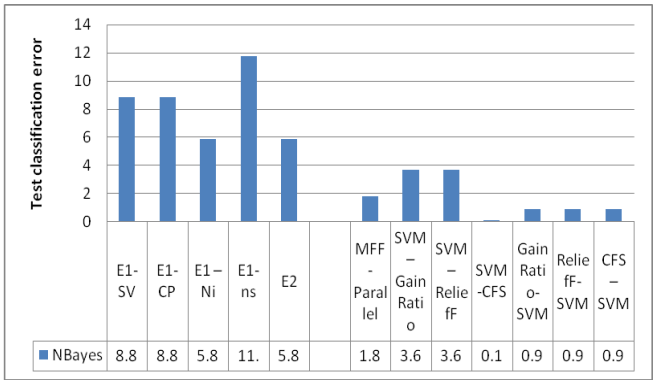

Figure 2: Comparison of test classification error between proposed MFF-EC method and ensemble of filters in existing method (Alonso-Betanzos at el .2014): Naive Bayes Classifier

V. DISCUSSION AND CONCLUSION

The simulation results shows that, both approaches of MFF-EC outperform the ensemble of classifiers in terms of average test classification value with minimum which is less than average of individual of classifier .In addition to that, MFF-parallel has been performing well than other six methods in terms of accuracy, sensitivity and specificity with average values respectively 98.56, 98.87 and 99.1.

The overall simulation results of these seven methods conclude that MFF-EC algorithm can effectively reduce the dimension of leukemia micro array data and select a small informative gene subset for classification with low classification error. Even though MFF-EC has fast computability and ability of dealing with high dimensionality data set, fails to take into account interaction with classifier as wrapper and embedded methods.

Improvement of the proposed MFF- EC method could be taken as future works. So, MFF-sequential method can be extended into other sequential filter arrangements with four filter method. Here we only discussed sequential combination of two filters with time limitations. In addition to that, this method can be extended into more general high dimensional data available area which highly requires feature selection.

Table 3: Comparison of test classification error between proposed MFF –EC method and existing multi filter approach (Alonso-Betanzos at el .2014).

	Ref.	Ibk	NB
E1-SV	Alonso-Betanzos at el .2014	14.7	8.82
E1-CP		14.7	8.82
E1 –Ni		**5.88**	**5.88**
E1-ns		11.76	11.76
E2		14.71	5.88
MFF-Parallel	Proposed MFF-EC approaches	**4.43**	1.84
MFF –Sequential			
SVM –Gain Ratio		9.73	3.68
SVM –ReliefF		21.82	3.68
SVM-CFS		7.92	**0**
Gain Ratio-SVM		10.57	0.92
ReliefF-SVM		13.21	0.92
CFS –SVM		14.05	0.92

REFERENCES

Abdullah, S Alzaqebah M. Nazri M. Z. A and Shreem S. S. (April 2013).A Hybrid Feature Selection approach of ensemble multiple Filter methods and wrapper method for Improving the Classification Accuracy of Microarray Data Set.In*: International Journal of Computer Science and Information Technology & Security*, 3(2).

Abeel, T. Dupont, P. Helleputte, T. Peer, Y.V. and Saeys, Y. (2010).Robust Biomarker Identification for Cancer Diagnosis with Ensemble Feature Selection Methods.*Bioinformatics*, 26(3). pp. 392-398.

Alonso-Betanzos, A. Bolón-Canedo, V. and Sánchez-Maroño, N. (2014).Data classification using ensemble of filters. *Neurocomputing*, 135(3), pp13-20.

Caporale, R.Huerta, E. Lopez, M. and Montiel, A. (August 2015) Hybrid Framework using Multiple-Filters and an Embedded Approach for an Efficient and Robust Selection and Classification of Microarray Data.In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 13(1), pp.1-1

Cybernetics, Part C (Applications and Reviews), 42(6), pp. 1590-1599.

Dittman, D. J. Khoshgoftaar, Napolitano ,A and Wald, T.M. (2013). Classication performance of rank aggregation techniques for ensemble gene selection. In: *C. Boonthum-Deneckeand G. M. Youngblood (Eds.*), FLAIRS Conference. AAAI Press

Etzioni, R. Feng, Z. Pepe, M.S. Potter, J.D. Thompson, M.L. Thornquist, M. Winget, M. and Y. Yasui (2001).Phases of Biomarker Development for Early Detection of Cancer.In*:J. Nat'l Cancer Inst*. 93(4), pp. 1054-1060.

Hussain, M. Khan, A. and Rathore, S. (Nov.-Dec. 1 2014).GECC: Gene Expression Based Ensemble Classification of Colon Samples. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 11(6), pp.1131-1145

Mao, K.Z. and Yang, F. (2011 July-Aug).Robust Feature Selection for Microarray Data Based on Multicriterion Fusion.In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(4), pp.1080-1092.

Mitra, S. and Ghosh, S. (Nov. 2012).Feature Selection and Clustering of Gene Expression Profiles Using Biological Knowledge. In: *IEEE Transactions on Systems, Man and*

Sasikala, S. (2014), Multi Filtration Feature Selection (MFFS) to improve discriminatory ability in clinical data set. *Applied Computing and Informatics*

Singer, B. Xiong, M. Yu, C. and Zhang, H. (2001).Recursive Partitioning for Tumor Classification with Gene Expression Microarray Data, In: *PNAS*, 98(12), pp. 6730-6735

Yukyee, L. and Yeungsam, H. (Jan.-March 2010).A Multiple-Filter-Multiple-Wrapper Approach to Gene Selection and Microarray Data Classification.In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(1), pp.108-117.

WHO<http://www.who.int/mediacentre/factsheets/fs297/en/>
25/1/2016

	Ibk	NavieBayes	AdaBoostM1	AdaBoostM1-Ibk	AdaBoostM1-NB	Bagging	Bagging-Ibk	Bagging-NB
MFF-Parallel	**4.43(0.03)**	1.84(3.88)	8.06(8.25)	**3.02(0.44)**	**0.12(2.91)**	26.91(6.28)	**3.57(0.38)**	2.74(3.54)
MFF -Sequential								
SVM -GainRatio	9.73(8.51)	3.68(6.43)	10.14(13.02)	3.63(8.85)	2.89(6.14)	25.92(5.05)	13.63(7.76)	2.92(4.60)
SVM -ReliefF	21.82(11.08)	3.68(7.75)	12.37(13.11)	20.76(12.38)	5.21(6.27)	26.54(6.94)	28.45(9.45)	3.93(5.69)
SVM-CFS	7.92(6.12)	**0.00(0.00)**	7.86(8.46)	8.15(14.40)	1.02(3.19)	**24.93(5.45)**	8.61(3.86)	**0.31(0.49)**
GainRatio-SVM	10.57(8.35)	0.92(2.91)	9.15(7.41)	5.96(8.76)	4.27(7.63)	25.80(5.31)	15.40(8.06)	1.47(1.58)
ReliefF-SVM	13.21(10.99)	0.92(2.91)	**5.29(5.73)**	8.41(8.31)	3.07(4.64)	26.86(5.85)	15.93(8.80)	1.28(1.45)
CFS -SVM	14.05(10.56)	0.92(2.91)	13.40(12.88)	9.80(11.90)	2.87(5.25)	31.34(6.40)	15.01(6.40)	2.36(2.68)
Individual Filter								
GainRatio	20.94(13.43)	9.10(10.61)	9.67(13.34)	17.15(12.75)	9.08(10.52)	25.17(5.02)	18.06(7.65)	9.36(8.84)
ReliefF	16.61(9.46)	9.98(10.13)	11.75(11.73)	15.34(9.91)	9.68(9.58)	25.78(6.93)	16.48(8.67)	10.34(9.08)
CFS	12.20(4.81)	**0.00(0.00)**	6.74(6.06)	11.49(8.63)	0.16(0.51)	30.91(6.20)	11.78(3.84)	0.45(0.77)
SVM	**4.43(0.03)**	1.84(3.88)	9.78(7.91)	3.05(0.45)	1.61(3.04)	25.66(6.73)	3.57(0.39)	2.74(3.56)

Table 4: Test classification error rate for leukemia data set