SPIRIT-ML: A Machine Learning Platform for Deriving Knowledge from Biomedical Datasets

Srisairam Achuthan(✉), Mike Chang, and Ajay Shah

Division of Research Informatics and Systems, Department of Information Sciences, City of Hope National Medical Center, Duarte, CA 91010, USA
sachuthan@coh.org

Abstract. SPIRIT-ML (Software Platform for Integrated Research Information and Transformation - Machine Learning) is a synergistic and flexible machine learning component of integrated research informatics platform, SPIRIT, being developed at City of Hope. SPIRIT-ML is being developed to analyze varied data analysis problems in biomedical and clinical datasets to further translational research. An interactive interface, broad spectrum of data driven learning models, multiple cross-validation techniques, visualization methods and reporting metrics constitute the platform.

Keywords: Machine learning · Translational research · Platform

1 Introduction

Machine learning algorithms have been applied to solve research problems encountered routinely in various biological and clinical settings. They have been applied to identify patient cohorts based on electronic medical record (EMR) data, identify malignant tumors based on image data, and for adverse drug surveillance based on publicly available databases, to name a few [2].

A biomedical dataset containing collection of patterns can be grouped into clusters based on similarity. For example, a cohort of cancer patients can be stratified into distinct clusters based on their demographic, biological and clinical characteristics. In solving a classification problem, we are interested in predicting the outcome (class label) of the dataset by building a model based on a training dataset. Similarly, malignant and benign tumors can be classified based on tumor characteristics from breast cancer patients [3,4]. Bayesian networks can help discover the dependent and independent variables [5] in high throughput molecular dataset consisting of genes and proteins identified in gene regulatory pathways.

Significant effort is spent when various machine learning methods are applied to biomedical and clinical problems using independent one-off deployment of computational pipeline. To address this problem we extended SPIRIT platform to include an interactive interface for applying a comprehensive set of

N. Ashish and J.-L. Ambite (Eds.): DILS 2015, LNBI 9162, pp. 240–250, 2015.
DOI: 10.1007/978-3-319-21843-4_19

machine learning methods. SPIRIT-ML allows normalization and binning of the input data, applies uniform data validation methods and creates reports that allow users to compare results across all the methods. SPIRIT-ML is specifically designed for extracting knowledge from biomedical datasets. SPIRIT-ML comes with a standard set of machine learning algorithms. Additional algorithms can be incorporated efficiently into SPIRIT-ML.

2 Methods

Increasingly, a large number of cloud based machine learning platforms are available for large scale data analysis and predictive analytics. These include, AWS Machine Learning platform [16], H2O [17], Apache Mahout [18] etc. WEKA [19] is an Open Source application that integrates several machine learning algorithms for data mining tasks. WEKA has been integrated with KNIME data pipelining tool to create data analysis pipeline. Our approach is similar, but singularly focused on biomedical data.

We utilize a commercial data pipelining and scientific informatics platform (Biovia's Pipeline Pilot, [21]) to integrate several machine learning algorithms from R, MATLAB, Huggin etc. One of the advantages of SPIRIT-ML is its ability to utilize a rich source of components in Pipeline Pilot. For example, Pipeline Pilot has an extensive collection of cheminformatics components. These components enable molecular similarity analysis, prediction of toxicology profiles of molecules and molecular database searching. Combing the machine learning algorithms with these cheminformatics algorithms can provide a powerful molecular classification application. Similarly, Pipeline Pilot provides access to a wide variety of computational protocols in bioinformatics. For example, Fig. 1 is a Pipeline Pilot protocol that is available to stratify Acute Lymphoblastic Leukemia (ALL) patients based on gene expression data [1]. Starting from the gene expression data from 32 ALL patients, using pairwise differential expression component along with intensity variation component one can extract the genes that meet the selection criteria. A PCA analysis of the corresponding microarrays is able to identify two clusters of patients as seen in Fig. 2. Pipeline Pilot protocols like this when integrated with machine learning algorithms in SPIRIT-ML increases its utility.

Pipeline Pilot can also be utilized to create machine learning applications for image analysis as well as text analysis. Image segmentation, morphology, transformations, image filtering and enhancement are some of areas within image analysis with multiple Pipeline Pilot protocols potentially available to SPIRIT-ML. For example, Per Object Thresholding component within image segmentation finds a different threshold for each biological cell in an array and successfully segments all arrays. Text analysis orientated Pipeline Pilot protocols can help crawl web pages, can create ontology files from different source formats, perform local searches (by indexing Pubmed for example) etc.

The entire SPIRIT-ML platform is built on top of Pipeline Pilot, a data pipeline software that provides a web interface for accepting user specified

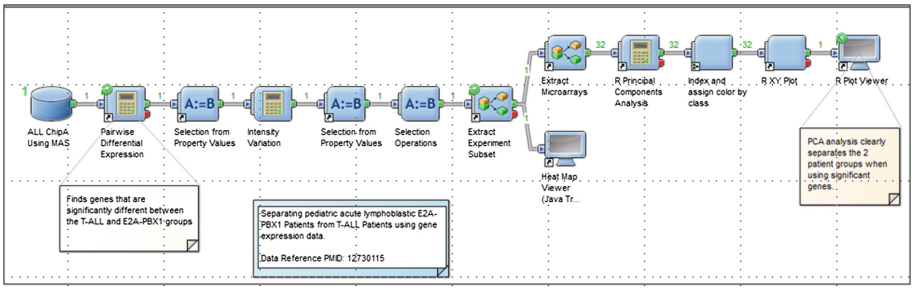

Fig. 1. A Pipeline Pilot protocol based on gene expression data to separate pediatric acute lymphoblastic E2A-PBX1 patients from T-ALL patients

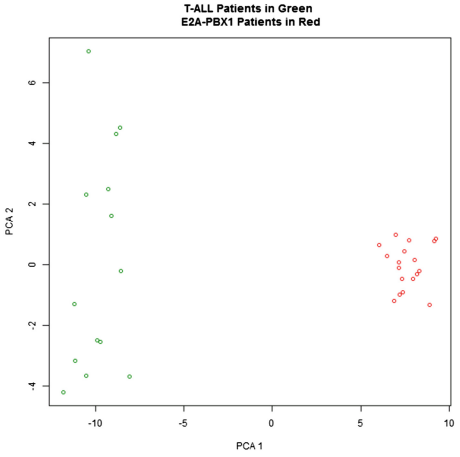

Fig. 2. A PCA of the gene expression data shows two clusters (T-ALL patients in green and E2A-PBX1 patients in red) (Color figure online)

options as well as displaying all the results obtained. Pipeline Pilot uses a data pipeling approach to handle and analyze research data. It uses a data flow framework to describe the processing of data. Algorithms written in R and MATLAB as well as external APIs provided by other scientific software applications like Hugin [20] can be integrated within the Pipeline Pilot environment. Using individual components, the entire data pre-processing, data analysis, visualization and web reporting is handled conveniently within Pipeline Pilot.

Figure 3 presents SPIRIT-ML approach to analyzing biomedical data sets. The data is grouped together as a matrix with columns containing features or attributes and rows containing observations or instances. The raw data is transformed via built-in options such as normalization and binning. Data attributes that are continuous are selected for normalization. Normalization scales the instances of each selected attribute to lie within unit range. The attributes that need to be binned can then be selected with user specified number of bins.

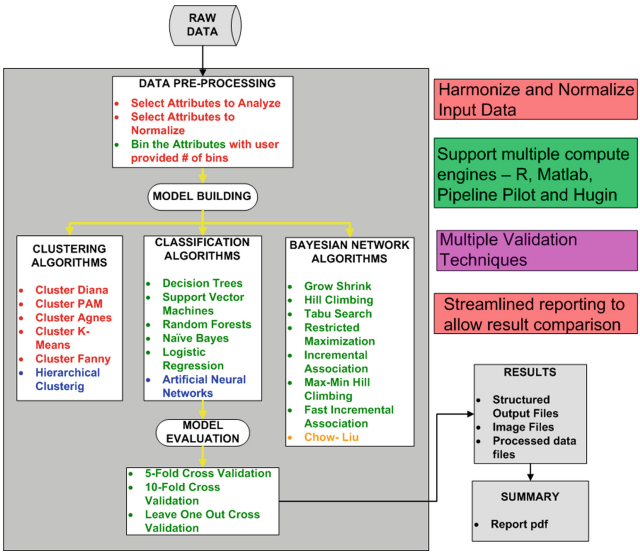

Fig. 3. SPIRIT-ML: The machine learning component of Software Platform for Integrated Research Information and Transformation. The texts in red color were implemented by using the R clustering algorithms implement in Pipeline Pilot version 9.1. The texts in green color were implemented by using different libraries in R. The texts in blue color were implanted using MATLAB (Color figure online).

SPIRIT-ML is a comprehensive framework for clustering, classifying, and deciphering relationships among covariates. It provides three types of data driven learning methods: unsupervised learning algorithms (i.e. clustering), supervised learning algorithms and Bayesian network models.

Clustering methods in SPIRIT-ML includes both hierarchical as well as non-hierarchical algorithms [6]. Agnes algorithm, an agglomerative hierarchical method, and DIANA algorithm, a divisive hierarchical method, are two hierarchical R clustering algorithms included as part of SPIRIT-ML. Non-hierarchical algorithms that are part of the clustering module of SPIRIT-ML include K-means [7], PAM and CLARA that can deal with large datasets. We have also implemented Cluster Fanny, a fuzzy clustering algorithm. Except, K-Means, all unsupervised learning algorithms are implemented using the cluster library in R. The K-Means algorithm is implemented using the stats library in R. Hierarchical algorithms with well known distance based metrics like Euclidean and Chebychev available in MATLAB are implemented in the SPIRIT-ML platform.

For classification problems, the transformed data is distributed with a fixed percentage utilized for training (usually 70 %), validating (15 %) and testing purposes (15 %) prior to developing the learning models. This is part of all R and MATLAB codes integrated within SPIRIT-ML. Decision tree [8] algorithm using the rpart library and C5.0 algorithm [9] using the C50 library were implemented

in R. These algorithms help convert the features into rules/decisions driven by the underlying data. Support Vector Machines (SVMs) [10] and Naïve Bayes algorithms [4] were implanted using the e1071 package in R. SVMs are useful to separate datasets using linear classifiers in a higher dimensional space. Independent features are best modeled using the Naïve Bayes classifier. Biologically inspired artificial neural networks [11] are able to approximate nonlinear functions and are referred to as nonlinear classifiers. MATLAB's neural network toolbox was utilized to implement this algorithm in SPIRIT-ML. Logistic regression i.e. multinomial log-linear model was implanted in R using the nnet library. These are best suited for classification problems where the class label is binary. Random Forests (RF [12]) implemented using the randomForests library in R is an ensemble of decision trees included in the suite of supervised learning algorithms.

SPIRIT-ML ranks features in decreasing order of their importance in building supervised learning models. For decision trees, the topmost node of the tree where maximum data instances are classified was used to identify the first rank. This process is iteratively applied on the remaining set of features. For SVMs, the root squared coefficients of the support vectors were used to rank all the features. For ANNs, the connection weights between the different layers were used to rank the features. For RF, the importance measure in the randomForest package in R was used to rank the features. For multinomial log-linear models, the exponential of logistic regression coefficients were used to rank the features.

For developing Bayesian networks from a given dataset, eight algorithms (Grow-Shrink, Hill-Climbing, Tabu Search, Restricted Maximization, Incremental Association, Max-Min Hill Climbing, Fast Incremental Association and Chow-Liu algorithms) have been implemented using the bnlearn library in R [13]. Hugin module within SPIRIT-ML can be utilized to create influence diagrams.

Multiple cross-validation methods [14] such as 5-fold, 10-fold as well as leave one out cross validation methods have been implemented. The outputs of the unsupervised models are clusters that may be visualized either as dendrograms or cluster plots. The performance of the supervised learning models are visualized using the Receiver Operating Characteristic curves (ROC). The results are summarized and made available as a pdf file. This file includes a side by side comparison of all the supervised learning model results obtained when solving a classification problem, the visualizations obtained from unsupervised learning models obtained when solving clustering problems and learning diagrams when constructing Bayesian network models.

3 Results

3.1 Predictive Model Building: Use Case 1

Fine needle aspiration (FNA) cytology characteristics of tumor cells differ between benign and malignant samples from breast cancer patients. Deterministic features measured by Dr. Wolberg and colleagues at University of Wisconsin

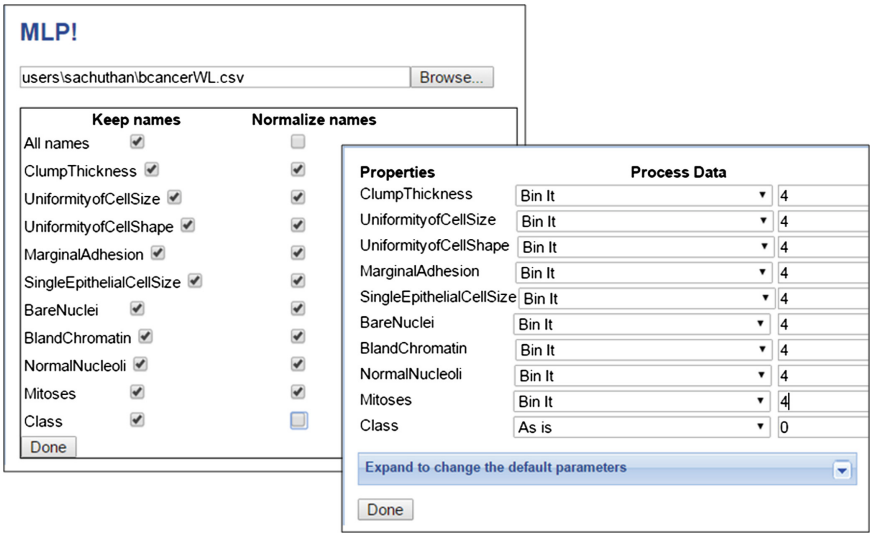

Fig. 4. Normalization and binning raw data

Hospitals from digitized image of aspirated cells [3,4] can be used to predict tumor malignancy. The dataset has been archived at UCI's machine learning repository and is referred to as the Wisconsin Breast Cancer dataset. This dataset contains 699 instances with each instance represented by a sample ID and 9 other numeric features. Supervised learning (classification) algorithms in SPIRIT-ML were utilized to predict tumor malignancy.

Figure 4 displays the features that were analyzed in the Wisconsin Breast Cancer dataset. The features assumed numeric values in the range of 1–10. The class labels were Benign as well as Malignant. In 16 instances, one of the nine features (BareNuclei) analyzed was missing a value. A random value between 1 and 10 was assigned for these instances. Since the nine features considered were continuous with values ranging between 1 and 10, we chose to normalize all of them. Each feature was then binned (4 bins with uniform width).

If the tab that indicates *Expand to change the default parameters* is selected, a drop down appears as a list of parameters (Fig. 5) for supervised learning models with editable values. For example, DT_minsplit, DT_CP and DT_minbucket refer to the minsplit, CP(Complexity Parameter) and minbucket parameters, respectively, expected by control option within the *rpart* command for the decision tree algorithm in R.

For the supervised learning task, SPIRIT-ML provides the results for all the algorithms side by side (Fig. 6). The first table compares the Accuracy of the seven supervised learning models based on the training dataset which in this case comprised of 70 % of the entire dataset. The second table in Fig. 6 compares the Precision, Recall, F_Measure and Specificity [15] across all the models. The final table in Fig. 6 ranks the top five features that led to the classification based

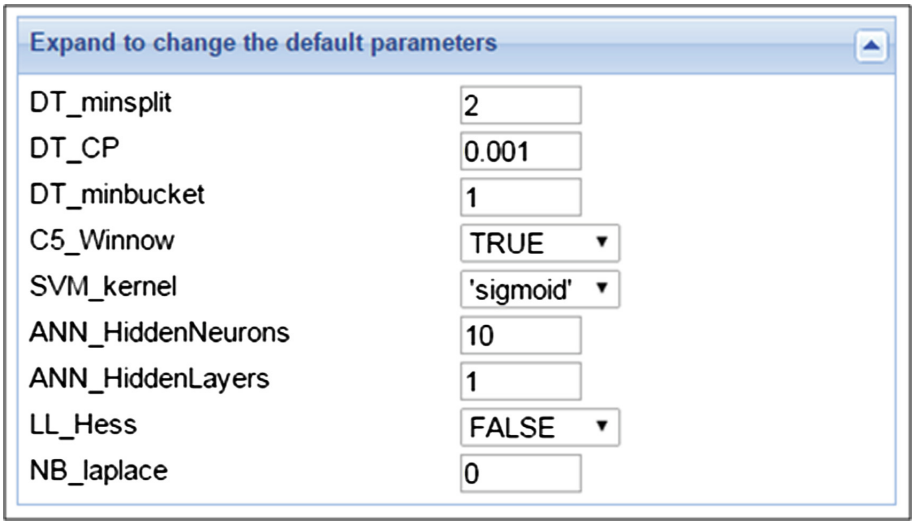

Fig. 5. Default parameters that can be modified

Prediction Accuracy with Training Data (Static Learning)

Data_Type	Decision Trees	C5 Trees	Support Vector Machines	Artificial Neural Networks	Random Forests	Log-Linear_Model	Naive Bayes
Correctly classified instances	488	475	472	478	339	475	473
Accuracy	99.796	97.137	96.524	98	69.325	97.137	96.728

Performance Comparison with Training Data

Data_Type	Decision Trees	C5Trees	Support Vector Machines	Artificial Neural Networks	Random Forests	Log-linear Model	Naive Bayes
Precision	1	0.98	0.98	0.97	0.68	0.98	0.98
Recall	1	0.98	0.96	0.99	1	0.98	0.97
F_Measure	1	0.98	0.97	0.98	0.81	0.98	0.97
Specificity	0.99	0.96	0.97	0.95	0.11	0.96	0.97

Top 5 Relevant Features that led to Classification

Decision Trees	Support Vector Machines	Artificial Neural Networks	Random Forests	Log-linear Model
BlandChromatin	Mitoses	Mitoses	BlandChromatin	UniformityofCellSize
UniformityofCellSize	MarginalAdhesion	ClumpThickness	ClumpThickness	NormalNucleoli
UniformityofCellShape	UniformityofCellShape	MarginalAdhesion	BareNuclei	SingleEpithelialCellSize
BareNuclei	SingleEpithelialCellSize	BareNuclei	UniformityofCellSize	BlandChromatin
SingleEpithelialCellSize	BlandChromatin	BlandChromatin	UniformityofCellShape	MarginalAdhesion

Fig. 6. Accuracy, Performance Measures and Feature Ranking based on training data

once on the training data. To combine the results of feature ranking we adopt the consensus polling method to determine the most important features. In our use case they were: BlandChromatin, UniformityofCellSize, ClumpThickness, MarginalAdhesion, UniformityofCellShape and BareNuclei.

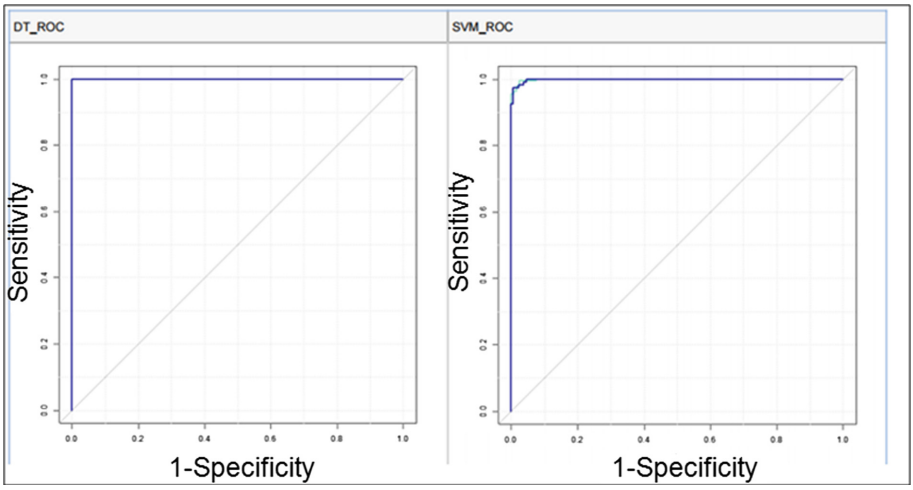

Fig. 7. ROC based on training data for two supervised learning algorithms

Prediction Accuracy with Test Data (Static Learning)							
Data_Type	Decision Trees	C5Trees	Support Vector Machines	Artificial Neural Networks	Random Forests	Log-linear Model	Naive Bayes
Correctly classified instances	101	100	100	100	71	99	101
Accuracy	96.19	95.238	95.238	95	67.619	94.286	96.19

Performance Comparison with Test Data							
Data_Type	Decision Trees	C5Trees	Support Vector Machines	Artificial Neural Networks	Random Forests	Log-linear Model	Naive Bayes
Precision	0.96	0.95	0.97	0.97	0.67	0.94	0.97
Recall	0.99	0.99	0.96	0.96	1	0.97	0.97
F_Measure	0.97	0.97	0.96	0.96	0.8	0.96	0.97
Specificity	0.91	0.89	0.94	0.94	0.03	0.89	0.94

Fig. 8. Accuracy, Performance Measures and Feature Ranking based on test data

The Receiver Operating Characteristic (ROC) curves based on the training data for two of the supervised learning models are shown in Fig. 7. They indicate that for the training data these models are quite good. To test these models, we used the test data against the models we created. The results for all the models are shown in Fig. 8. We found that the accuracy and performance on test data are greater than 94 % for all models except that obtained by Random Forests algorithm.

3.2 Clustering Dataset: Use Case 2

Identifying relevant patient characteristics in the case of complex diseases such as diabetes, cancer and dementia is quite challenging. Patient demographics, diagnosis and procedure information are usually captured in coded format within

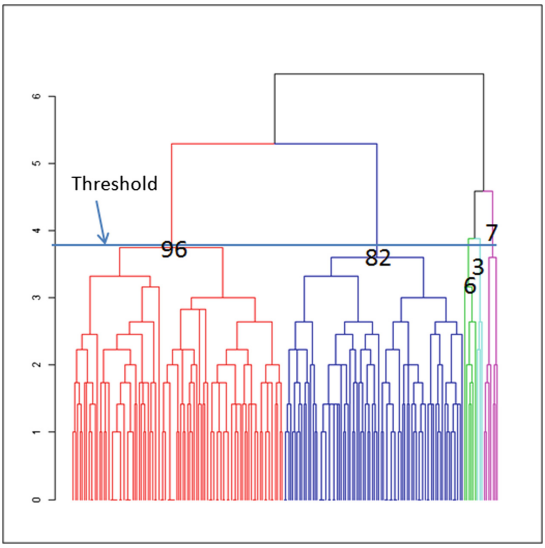

Fig. 9. Dendrogram depicting 5 clusters using Cluster DIANA

clinical databases. Patterns within the coded fields may reveal clinical characteristics across patients that would be difficult to determine manually. To automate this process, we developed a use case based on coded features derived from MIMIC II dataset (http://mimic.physionet.org/), a publicly available clinical database, using SPIRIT-ML.

Type II diabetic patients with certain types of cancer (Liver, Pancreatic, Uterus, Colon, Bladder, Breast, Kidney, Esophageal and Ovarian) within MIMIC II dataset were clustered using the clustering algorithms in SPIRIT-ML to reveal clinical characteristics across patients. A total of 194 patients (instances) with ten features (Age, Gender, Ethnicity, BMI, Congestive Heart Failure Yes/No, Cardiac Arrthymias Yes/No, Hypertension Yes/No, Chronic Pulmonary Yes/No, Renal Failure Yes/No and Liver Disease Yes/No) were analyzed.

Figure 9 is a dendrogram visualization of the five clusters obtained by Cluster DIANA algorithm, one of the six clustering algorithms implemented in SPIRIT-ML. The threshold decides the number of clusters identified. In this case, there are two major clusters and three minor clusters. The numbers in Fig. 10 indicate the number of patients in each cluster. The characteristics that decide the cluster membership for each patient can be determined by converting the clustering problem into a classification problem where the cluster membership is taken to be the class label. Figure 8 depicts the cluster plot for the five clusters. These plots are helpful in visualizing where the individual clusters lie in relation to other clusters. The clusters with 3 and 7 patients seemed to be within the cluster with 96 patients. The cluster with 6 patients overlaps with the two major clusters. This plot suggests that in reality there are only two main patient clusters.

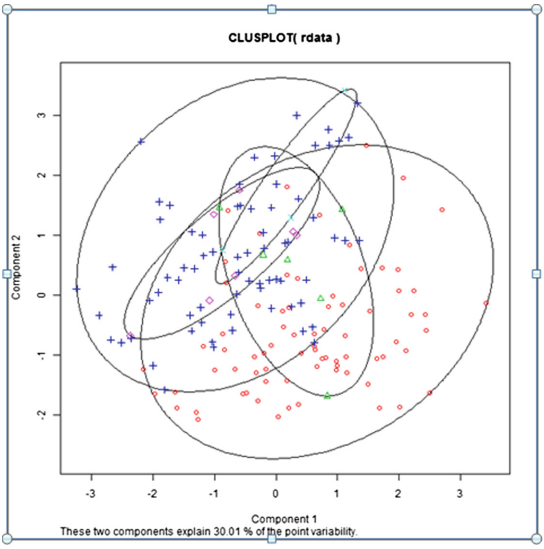

Fig. 10. Cluster plot using cluster DIANA

4 Future Work

The future development of platform will include vertical and horizontal integration to enable integrated research informatics. The horizontal integration currently planned include integration of SPIRIT-ML with the biomedical and clinical natural language processing component of SPIRIT (SPIRIT-NLP), image analysis and genomics computational pipelines being implemented at COH as part of SPIRIT platform.

The vertical integration of SPIRIT-ML with the n-tier SPIRIT platform is also being planned. This integration will include application integration using FUSION middleware, web services, common user interface components shared amongst other SPIRIT applications.

5 Conclusions

SPIRIT-ML is a functional machine learning platform that is used to discover and reveal patterns in biomedical datasets. SPIRIT-ML provides the following features: (a) Normalization and harmonization of input data (b) Clustering, classification and Bayesian network algorithms for deciphering relationships within a dataset (c) Various Validation methods (d) Integrated reporting system for comparative analysis of results. The underlying design of the platform is flexible enough to include machine learning models of choice, and facilitates comparison of results obtained by each model side by side. With the aid of SPIRIT-ML, the needs of multiple translational research projects that require data driven knowledge extraction can be addressed. We intend SPIRIT-ML to be an open source platform so that machine learning methods developed in other packages such as WEKA can be incorporated with minimal effort via Web Services.

Acknowledgments. The authors would like to thank Dr. Joyce Niland, Dr. Haiqing Li and Dr. Weizhong Zhu for their input and feedback.

References

1. Ross, M.E., Zhou, X., et al.: Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. Blood **102**(8), 2951–2959 (2003)
2. Cleophas, T.J., Zwinderman, A.H.: Machine Learning in Medicine. Springer, Netherlands (2013)
3. Wolberg, W.H., Mangasarian, O.L.: Multisurface method of pattern separation for medical diagnosis applied to breast cytology. PNAS **87**, 9193–9196 (1990)
4. Zhang, J.: Selecting typical instances in instance-based learning. In: Proceedings of the Ninth International Machine Learning Conference, Aberdeen, Scotland, pp. 470–479. Morgan Kaufmann (1992)
5. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian network classifiers. Mach. Learn. **29**(2–3), 131–163 (1997)
6. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York (1990)
7. Hartigan, J.A.: Clustering Algorithms. Wiley, New York (1975)
8. Breiman, L., Freidman, J., Stone, C.J., Olshen, R.A.: Classification and Regression Trees. Chapman and Hall/CRC, Boca Raton (1984)
9. Kuhn, M., Johnson, K.: Applied Predictive Modeling. Springer, New York (2013)
10. Cortes, C., Vapnik, V.: Support-vector network. Mach. Learn. **20**, 1–25 (1995)
11. Werbos, P.J.: Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. Ph.D. thesis, Harvard University (1974)
12. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
13. Nagarajan, R., Scutari, M., Lebre, S.: Bayesian Networks in R: with Applications in Systems Biology. Springer, New York (2013)
14. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelliegence, vol. 2(12), pp. 1137–1143 (1995)
15. Powers, D.M.W.: Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. J. Mach. Learn. Technol. **2**(1), 37–63 (2011)
16. Amazon Web Services Machine Learning. http://aws.amazon.com/machine-learning/
17. H_2O - the open source predictive analytics platform. http://0xdata.com/product/
18. The Apache Mahout. http://mahout.apache.org/
19. Waikato Environment for Knowledge Analysis (WEKA). http://www.cs.waikato.ac.nz/ml/weka/
20. Pipeline Pilot platform. http://accelrys.com/products/pipeline-pilot/
21. Hugin, the decision support tool. http://www.hugin.com/productsservices/products/academic/researcher