www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 4 Issue 9 Sep 2015, Page No. 14244-14247

Efficient Recognition Of Survival Rate In Bone Marrow Records Using Non-Matrix Factorization Algorithm

S.Vinothini[1], S.C.Punitha[2].
[1]Research scholar , PSGR Krishnammal college for women ,
Coimbatore, Tamilnadu.
Svino38@gmail.com

[2]S.C.Punitha, Dept of computer science,
PSGR Krishnammal college for women,
Coimbatore.

Abstract: ***Patients undergoing a bone marrow stem cell transplant face various risk factors, stem cell is a procedure to replace the damaged or destroyed bone marrow with healthy bone marrow. Stem cells are immature cells in the bone marrow that gives rise to all of blood cells. Totally 2000 patients records are collected also split into training and test data of 768 records with 18 different attributes. Non-matrix factorization algorithm are used to find the missing values, for handling the more attributes and more data set of different patients with excluding of errors and missing values. The data set focused in targeted information extraction and investigative analysis along with useful patterns. Various classification algorithms like SVM, RF, and NN are trained on predicting the survival of each patient depends on their preoperative measurements along with highest prominence. Non-matrix algorithm increases the accuracy of prediction result.***

Keywords: hematopoietic stem cell, chemo-sensitive, bone marrow, SVM, RF, NN.

1. Introduction

The blood cells in human body start out as immature cells called hematopoietic stem cells. Stem cells generally live in the bone marrow, where it divides to make new blood cells. Once the blood cells mature, it will leave the bone marrow and enter the bloodstream of the body is called peripheral blood stem cells.

Stem cell transplants are used to reinstate the stem cells when the bone marrow has been destroyed by disease, chemotherapy, or radiation. Depending on the source of the stem cells, the transplant procedure may be called as bone marrow transplant, peripheral blood stem cell transplant, or cord blood transplant. These are called hematopoietic stem cell transplants.

Nowadays hundreds of thousands of patients have had stem cell transplants. There are three possible sources of stem cells are used for transplants; Bone marrow, bloodstream and Umbilical cord blood from newborns.

Typically the stem cell transplant is used for destroy the cancer cells by using radiation therapy, this treatment also kills the stem cells in the bone marrow. Shortly after treatment stem cells are given to replace those that were destroyed. Then the stem cells are settle in the bone marrow and begin to grow and make healthy blood cells. This procedure is called engraftment. There are 3 types of transplants based on who gives the stem cells.

Autologous Stem Cell Transplantation -The stem cells are taken from the patients itself.

Allogeneic Stem Cell Transplantation -The stem cells are taken from a matched related or unrelated donor.

Syngeneic stem cell transplants -The stem cells are taken from twin or triplet of the identical sibling.

In the existing scenario, the collaborative filtering techniques and classification algorithms were used to classify the patients who have undergone stem cell transplant with high odds of survival and also keeping track of data about the donors within the family and outside the family which has a direct impact in the prioritization of resources. The probabilistic Principal Component Analysis (PCA), probabilistic matrix factorization and Robust PCA are used to handle the missing values. Even machine-learning techniques are used to identify the patients correctly; it has issue with the explicit modeling of the binary properties of dissected features. Also while increasing the number of attributes and patients records it does not suitable for predicting the result. Hence the go for proposed scenario, to overcome these issues.

The remaining paper is organized as follows; in section II defines the related of this study. Section III deals with methodology used in this proposed work. Section IV gives the brief detail of the proposed work with experimental results of various classification algorithms. Finally section V gives the conclusion and future work.

2. Related Work

Babak Taati et al [1], Analyze data from past transplants could enhance the understanding of the factors influencing success. Records up to 120 measurements per transplant procedure from 1751 patients undergoing BMT were collected (Shariati Hospital). Collaborative filtering techniques allowed the processing of highly sparse records with 22.3% missing values. Ten-fold cross-validation was used to evaluate the performance of various classification algorithms trained on predicting the survival status. Modest accuracy levels were obtained in predicting the survival status (AUC = 0.69). More importantly, however, operations that had the highest chances of success were shown to be identifiable with high accuracy, e.g., 92% or 97% when identifying 74 or 31 recipients, respectively.

Jayshree et al [2], uses Support Vector Machine to classify the patients who have undergone stem cell transplant with high odds of survival and also keeping track of information about the donors within the family and outside the family which has a direct impact in the prioritization of resources. Classification of this information is useful to create the need for a global perspective for all cell, tissue, and organ transplants and to reveal statistical structure with potential implications in evidence-based prioritization of resources.

Baron et al [3], linear discriminant analysis was used to identify “stronger alloresponders”, that is, donors who are more likely to elicit GvHD. The study aimed to predict any GvHD in the patient post-transplant, and was able to identify stronger alloresponders with up to 80% accuracy comparing 17 genes and four gene pairs. The gene profile expressions if used require extensive manual analysis and potential costs, which is not suitable for clinical applications

Petersdorf et al [4], identifying donors whose cell transplantation could result in GvHD was also investigated using logistic regression to associate haplotype mismatches with grades III–IV aGvHD. The authors examined three traits from the recipients and five traits from the donors, and found that the haplotype mismatches statistically corresponded to increased risk of severe aGvHD.

Taati et al [5], present the information about stem cell transplant records are often merged from different resources so the data is not available for analysis. The ten-fold cross validation technique is used basically to evaluate the performance of various classification techniques or algorithms to predict the results related to research.

Tomblyn et al [6], gives information about the complications during the stem cell transplant and provides a solution to reduce the infections. In allogeneic transplant, where the stem cells are taken from another person there is a risk that the patient may get infected. This can lead to graft-versus-host disease which causes the patient to suffer more after the stem cell transplant.

3. Methodology

3.1Proposed Work

The above figure (fig1) shows the proposed frame work. In the proposed work Non-negative matrix factorization (NMF) method is used for handling missing values to find parts-based linear representations of non-negative data. It is more robustly used for improving the decompositions as well as explicitly incorporating the notion of sparseness.

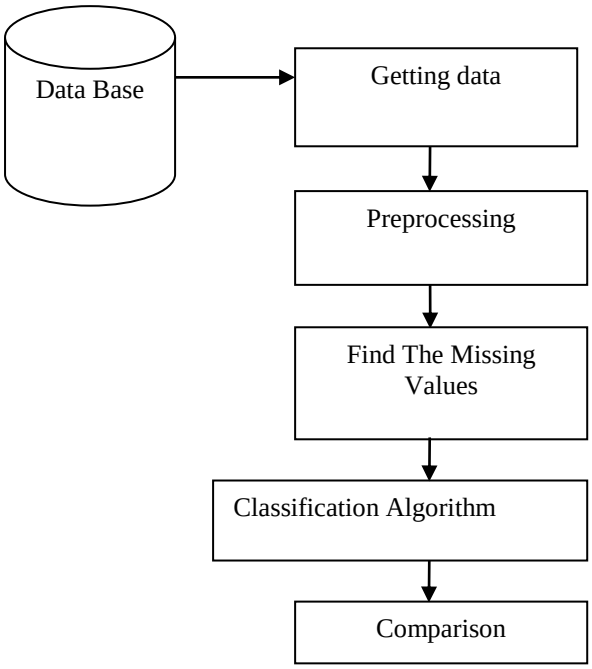

Fig1: Proposed framework

The NMF is used for selecting the most important features from the high dimensional dataset. This is also composed of the

two associated components of irrelevant as well as redundant feature elimination. It reduces the error values significantly and handles the missing information efficiently. It provides optimal decisions as to the prioritization of surgical procedures and the allocation of other resources to save the highest number of lives possible. There are three machine learning algorithms used for classification such as Support Vector Machine (SVM), Random Forest (RF) and Neural Network (NN).

3.2. Non-negative matrix factorization (NMF)

Non-negative matrix factorization (NMF) given a non-negative matrix V, find non-negative matrix factors Wand H such that:

$$V \sim WH \quad (1)$$

NMF can be applied to the statistical analysis of multivariate data in the following manner. Given a set of of multivariate n-dimensional data vectors, the vectors are placed in the columns of an n x m matrix V where m is the number of examples in the data set. This matrix is then approximately factorized into an n x r matrix Wand r x m matrix H [7].

Usually r is neither chosen to be smaller than nor m, so that Wand H are smaller than the original matrix V. This results in a compressed version of the original data matrix.

Cost functions

To find an approximate factorization V ~ WH, we first need to define cost functions that quantify the quality of the approximation. Such a cost function can be constructed using some measure of distance between two non-negative matrices A and B. One useful measure is simply the square of the Euclidean distance between A and B

$$\|A - B\|^2 = \sum_{ij} \left(A_{ij} - B_{ij}\right)^2 \quad (2)$$

This is lower bounded by zero, and clearly vanishes if and only if A = B.

3.3 Classification Algorithm

3.3.1 Random Forest (RF)

Random forest (or random forests) is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees. The method combines Breiman's "bagging" idea and the random selection of features.

Random forest algorithm

Each tree is constructed using the following algorithm:

1. Let the number of training cases be N, and the number of variables in the classifier be M.
2. We are told the number m of input variables to be used to determine the decision at a node of the tree; m should be much less than M.
3. Choose a training set for this tree by choosing n times with replacement from all N available training cases (i.e. take a bootstrap sample). Use the rest of the cases to estimate the error of the tree, by predicting their classes.
4. For each node of the tree, randomly choose m variables on which to base the decision at that node. Calculate the best split based on these m variables in the training set.
5. Each tree is fully grown and not pruned (as may be done in constructing a normal tree classifier).

For prediction a new sample is pushed down the tree. It is assigned the label of the training sample in the terminal node it ends up in. This procedure is iterated over all trees in the ensemble, and the average vote of all trees is reported as random forest prediction.

3.3.2 Support Vector Machine (SVM)

Support Vector Machines (SVMs), a classification method based on maximum margin linear discriminants, that is, the goal is to find the optimal hyperplane that maximizes the gap or margin between the classes. Further, use the kernel trick to find the optimal nonlinear decision boundary between classes, which corresponds to a hyper plane in some high-dimensional “nonlinear” space.

Algorithm

1. Choose a kernel function
2. Choose a value for C
3. Solve the quadratic programming problem (many software packages available)
4. Construct the discriminant function from the support vectors.

3.3.3 NN algorithm

Neural networks were modeled after the cognitive processes of the brain. That is capable of predicting new observations from existing observations. A neural network consists of interconnected processing elements also called units, nodes, or neurons. The neurons within the network work together, in parallel, to produce an output function. Since the computation is performed by the collective neurons, a neural network can

still produce the output function even if some of the individual neurons are malfunctioning.

4. Experiment and results

4.1Results

From the below table (table 1), its proven that the NN works superior among these three and provides high accuracy results in the prediction. From the experimental it observes that the proposed method of NN algorithm shows highest accuracy, precision and recall values for more accurate prediction of bone marrow transplant.

Table1. Performance evaluation measures for various classification algorithms.

Algo*	Recall	Precision	Specificity	Acc*
RF	0.77|0.82	0.79|0.80	0.80|0.79	0.79
SVM	0.80|0.83	0.85|0.77	0.77|0.85	0.81
NN	0.90|0.92	0.91|0.91	0.91|0.91	0.91

*Algo-Algorithm, Acc-Accuracy

The following figure (fig2) shows the comparison chart for classification accuracy of various algorithms.

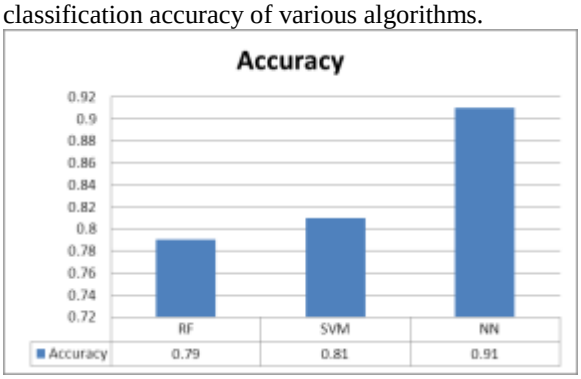

Fig 2. Comparison chart for classification accuracy.

4.1Conclusion

Records and extents from the past hematopoietic stem cell transplant procedures were analyzed to investigate the possibility of predicting the survival status of each patient based on their pre-operative information and test results. The experimental results proven that non-negative matrix factorization algorithm increases the accuracy of prediction results as well as performance of the proposed research improved significantly.

Future work includes incorporating a generative model on the distribution of missing values into the prediction process, and also the collection of further records to enrich the dataset for further analysis. In addition to that future selection method is used to finding out best future for yield more accuracy.

References

[1].Babak Taati, Jasper Snoek, Dionne Aleman, "Data Mining in Bone Marrow Transplant Records to Identify Patients with High Odds of Survival", IEEE Journal Of Biomedical and Health Informatics, VOL. 18, NO. 1, JANUARY 2014, pp: 21-27.

[2]. Ms. Jayshree S. Raju, Mr. Prafulla L. Mehar, "A Review Paper on Classification of Stem Cell Transplant to Identify the High Survival Rate", International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC), Volume: 3 Issue: 4, April 2015, pp: 1918 – 1920.

[3]. C. Baron, R. Somogyi, L. D. Greller, V. Rineau, P. Wilkinson et al., "Prediction of graft-versus-host disease in humans by donor gene-expression profiling," PLoS Med., vol. 4, no. 3, pp. 69–83, 2007.

[4]. E.W. Petersdorf, M. Malkki, T. A. Gooley, P. J.Martin, and Z. Guo,"MHC haplotype matching for unrelated hematopoietic cell transplantation," PLoS Med., vol. 4, no. 3, pp. 59–68, 2007.

[5]. B. Taati, J. Snoek, D. Aleman, and A. Ghavamzadeh, "Data mining in bone marrow transplant records to identify Patients with high odds of survival", IEEE journal, vol.18,no.1, 2014.

[6]. M. Tomblyn, T. Chiller, H. Einsele, R. Gress, K. Sepkowitz, "Guidelines for Preventing Infectious Complications among Hematopoietic Cell Transplantation Recipients: A Global Perspective",ASBMT, 2009

[7]. Daniel D. Lee, H. Sebastian Seung, "Algorithms for Non-Negative Factorization", Proceedings of the Conference on Neural Information Processing Systems 9, 515- 521.

[8]. Paatero, P & Tapper, U (1997). Least squares formulation of robust non-negative factor analysis. Chemometr. Intell. Lab. 37, 23- 35.

[9]. Andy Liaw and Matthew Wiener, "Classification and Regression by randomForest", R News, Vol. 2/3, December 2002, ISSN 1609-3631, pp:18-22.

[10]. Cortes, Corinna, Vapnik and Vladimir N, "Support Vector Networks, Machine Learning".

[11]. F.Fogelman Soulie, "Neural Network Architecture and Algorithms:A-Prespective", in T.Kohonen, K.Makisara, O.Simula and J.Kangas (eds) Artifical Neural Networks:Proc.of ICANN, Elsevier North-Holland.