International Journal of Scientific Engineering and Applied Science (IJSEAS) - Volume-1, Issue-8,November 2015
ISSN: 2395-3470
www.ijseas.com

Efficient Identification Of Patients In Bone Marrow Transplant Records Using Non- Matrix Factorization Algorithm

S.Vinothini[1], S.C.Punitha[2]
[1]PSGR Krishnammal Collage For Women, Coimbatore, India
[2]PSGR Krishnammal Collage For Women, Coimbatore, India

Abstract

A bone marrow transplant is a procedure to replace the unhealthy bone marrow to healthy bone marrow. Totally 2000 patient records are collected. A patient suffering hematopoietic stem cell transplant faces several risk factors. It has become the typical of care for habitual or attained disorders of the hematopoietic structure or with chemo-sensitive, radiosensitive or immune sensitive malignancies. The Non-Matrix Factorization algorithm used to finding missing values, in bone marrow records, and PSO algorithm used to select the important features, and it used to get the better accuracy compare to others. After apply the Machine learning algorithms like SVM, NN, RF are applied to predicting the survival of each patients depends on their preoperative measurements along with highest prominence .NMF and PSO algorithm increase the accuracy of prediction result.

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

1. Introduction

A bone marrow transplant is a procedure performed to replace bone marrow that has been damaged or destroyed by disease or chemotherapy. This procedure involves transplanting blood stem cells, which travel to the bone marrow where they produce new blood cells and promote growth of new marrow. Bone marrow is the soft, sponge-like material found inside bones. It contains immature cells known as hematopoietic or blood-forming stem cells. Hematopoietic stem cells divide into one of three types of blood cells, white blood cells, which are integral in fighting infection, red blood cells, which carry oxygen and nutrients through the body, and platelets, which are responsible for the formation of clots.

A bone marrow transplant replaces your damaged stem cells with healthy cells. This helps your body make enough white blood cells, platelets, or red blood cells to avoid infections, bleeding disorders, or anemia. Healthy stem cells can come from a donor, or they can come from your own body. In such cases, stem cells can be harvested before a person begins chemotherapy or radiation treatment. Those healthy cells are then stored and used in transplantation. Some reasons for a bone marrow transplant include:

- **aplastic anemia** is to disorder where the marrow stops making new blood cells.
- **cancers** that affect the marrow such as a leukemia, lymphoma, and multiple myeloma and other type of cancers.
- **damaged bone marrow** due to chemotherapy, To change the damaged bone marrow to healthy marrow.
- **congenital neutropenia** an inherited disorder which causes recurring infections.
- **sickle cell anemia** an inherited blood disorder that causes misshapen red blood cells.
- **thalassemia** an inherited blood disorder where the body doesn't make enough red blood cells.

Machine learning Algorithm

Machine learning is a subfield of computer science (CS) and artificial intelligence (AI) that deals with the construction and study of systems

that can learn from data, rather than follow only explicitly programmed instructions.

Neural Network

A neural network is a type of computational model which is able to solve multi problems in various fields. It processes the information in a similar way as the human brain concept processing the information. Basically, neural network consists of large processing elements called neurons working together to perform specific tasks. As in the human brain, there are thousands of dendrites which contain information signals. They transmitted the signals to the axon in the form of electrical spikes. The axon then sends the signals to another dendrites causing to a synapse.

Support Vector Machine

Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. A support vector machine constructs a hyper plane or set of hyper planes in a high-dimensional space, which can be used for classification, regression or other tasks.

Random forest

One of the most popular methods or frameworks used by data scientists at the Rose Data Science Professional Practice Group is Random Forests. The Random Forests algorithm is one of the best among classification algorithms - able to classify large amounts of data with accuracy.

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.

3. Methodology

3.1Proposed Framework

The NMF is used for selecting the most important features from the high dimensional dataset.

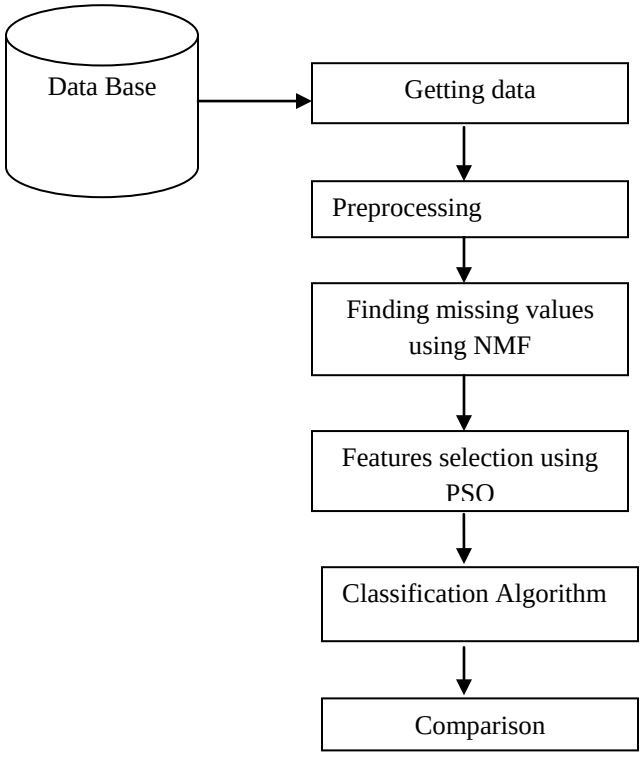

Fig1: Proposed framework

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.

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. 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. After PSO algorithm was used to select the important features.

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:

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

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

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

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

3.3 PSO Algorithm

A basic variant of the PSO algorithm works by having a population of candidate solutions called particles. These particles are moved around in the search-space according to a few simple formulae. The movements of the particles are guided by their own best known position in the search-space as well as the entire swarm's best known position. When improved positions are being discovered these will then come to guide the movements of the swarm. The process is repeated and by doing so it is hoped, but not guaranteed, that a satisfactory solution will eventually be discovered.

Formally, let f $\mathbb{R}^n \rightarrow \mathbb{R}$ be the cost function which must be minimized. The function takes a candidate solution as argument in the form of a vector of real numbers and produces a real number as output which indicates the objective function value of the given candidate solution. The gradient of f is not known. The goal is to find a solution $\mathbf{a}$ for which $f(\mathbf{a}) \leq f(\mathbf{b})$ for all $\mathbf{b}$ in the search-space, which would mean $\mathbf{a}$ is the global minimum. Maximization can be performed by considering the function $h = -f$ instead.

Let S be the number of particles in the swarm, each having a position $\mathbf{x}_i \in \mathbb{R}^n$ in the search-space and a velocity $\mathbf{v}_i \in \mathbb{R}^n$. Let $\mathbf{p}_i$ be the best known position of particle i and let $\mathbf{g}$ be the best known position of the entire swarm. A basic PSO algorithm is then.

- For each particle $i = 1, ..., S$ do:
  - Initialize the particle's position with a uniformly distributed random vector: $\mathbf{x}_i \sim U(\mathbf{b_{lo}}, \mathbf{b_{up}})$, where $\mathbf{b_{lo}}$ and $\mathbf{b_{up}}$ are the lower and upper boundaries of the search-space.
  - Initialize the particle's best known position to its initial position: $\mathbf{p}_i \leftarrow \mathbf{x}_i$
  - If $(f(\mathbf{p}_i) < f(\mathbf{g}))$ update the swarm's best known position: $\mathbf{g} \leftarrow \mathbf{p}_i$
  - Initialize the particle's velocity: $\mathbf{v}_i \sim U(-|\mathbf{b_{up}}-\mathbf{b_{lo}}|, |\mathbf{b_{up}}-\mathbf{b_{lo}}|)$
- Until a termination criterion is met (e.g. number of iterations performed, or a solution with adequate objective function value is found), repeat:
  - For each particle $i = 1, ..., S$ do: For each dimension $d = 1, ..., n$ do:
- Update the particle's position: $\mathbf{x}_i \leftarrow \mathbf{x}_i + \mathbf{v}_i$

$$\text{If } (f(\mathbf{x}_i) < f(\mathbf{p}_i)) \qquad (3)$$

- Update the particle's best known position: $\mathbf{p}_i \leftarrow \mathbf{x}_i$
- If $(f(\mathbf{p}_i) < f(\mathbf{g}))$ update the swarm's best known position: $\mathbf{g} \leftarrow \mathbf{p}_i$

Now $\mathbf{g}$ holds the best found solution.

3.4 Classification Algorithm

3.4.1Random 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.

3.4.2 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.

3.4.3 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. Experimental And Results

4.1. Results

From the below table (table 1), its proven that the NN works superior among these three and provides high accuracy results in the prediction.

Table1. Performance evaluation measures for various classification algorithms.

Algo *	Recall	Precision	Specificity	Acc *
RF	0.86|0.87	0.80|0.92	0.87|0.86	0.87
SVM	0.92|0.94	0.89|0.95	0.94|0.92	0.93
NN	0.95|0.97	0.95|0.97	0.97|0.95	0.96

*Algo-Algorithm, Acc-Accuracy

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.

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

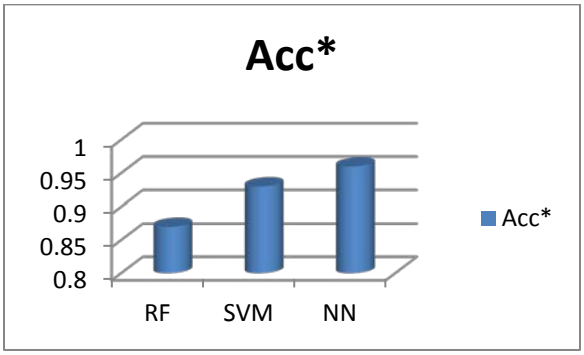

Fig 2. Comparison chart for classification accuracy.

5. Conclusions

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 the explicit modeling of the binary properties of dissected features into non- matrix factorization, incorporating a generative model on the distribution of missing values into the prediction process, and also the collection of further records (more patients and more attributes) to enrich the dataset for further analysis.

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] Mithun Das Gupta and Jing Xiao," Non-negative Matrix Factorization as a Feature Selection Tool for Maximum Margin Classifiers", in proc of IEEE transactions, 2011.

[10] Dingding Wang Tao Li and Chris Ding, "Weighted Feature Subset Non-Negative Matrix Factorization and its Applications to Document Understanding", in the proc of IEEE International Conference on Data Mining, 2010.

[11] Patrik O. Hoyer, "Non-negative Matrix Factorization with Sparseness Constraints", In the proc of Journal of Machine Learning Research, 2004.

[12] B. Taati, J. Snoek, D. M. Aleman, A. Mihailidis, and A. Ghavamzadeh, "Applying collaborative filtering techniques to data mining in bone marrow transplant records," in *Proc. 7th INFORMSWorkshop Data Mining Health Informat.*, Phoenix, AZ, USA, Oct. 2012.