RESEARCH ARTICLE OPEN ACCESS

Machine Learning Techniques for Bioinformatics and Computational Biology- A Survey

Babasaheb .S. Satpute [1], Dr. Raghav Yadav [2]
Department of Computer Science & I.T
SIET, SHUATS
Allahabad - India

ABSTRACT
There has been astronomical growth in the amount of biological data over the years and hence increased complexity and size of the databases storing those biological data. Hence we need most advanced and technically influential technology and tools to use for the analysis of those data. Machine learning field has emerged as of immense important and use in recent days. These days Machine learning algorithms are extensively in use for solving complex bioinformatics problems. The machine learning algorithms have also become very popular names in the Research Fraternity. Those algorithms have been used in large extent in bioinformatics applications. In this paper we present the review of some of the famous supervised classification and Clustering algorithms.
Keywords:— Machine learning, bioinformatics, supervised classification, clustering, classifier evaluation.

I. INTRODUCTION

In recent times huge amount of biological data is being generated due to lot of experiments and computations. In order to store such a huge data use of computers has become imperative. So lot of databases for storing, managing and manipulating such data have emerged in recent times. Traditional data analysis tools are also not enough, so researcher were need of some advanced technological tools and algorithms for analyzing such a vast data and hence Machine learning algorithms have emerged as an important tools for such kind of algorithms. The machine learning algorithms have become very popular names in the Research Fraternity. The basic idea supporting those algorithms is learning automatically from a dataset. Here we first make the algorithm learn or train it by using the training dataset and then we test the performance of those algorithms by using the test dataset. If the performance is up to the mark then we can use those algorithms for unknown dataset and unknown problems. The algorithms using above idea are called as supervised learning algorithms or supervised classifiers. The algorithms which are not trained by using a training dataset but can learn themselves are called unsupervised learning algorithms. Machine learning methods do lot of intensive computations and hence need lot of computational power but over the years computers and their computational power has been evolved to large extent hence machine learning are getting benefitted. Machine learning algorithms are basically the computer programs which make use of the dataset and past knowledge to solve the unknown problems. The main aim of this report is to present the review of the popular machine learning techniques.

Prompramote et al. has listed following few causes of large use of various machine learning algorithms in the field of computational biology and bioinformatics:

1. Experts are not generally ready to depict the variables they consider while evaluating a circumstance or while clarifying the principles they apply in ordinary practice. Machine learning can fill in as a profitable guide to separate the portrayal of the shrouded circumstance as far as those components and after that propose the tenets that better depict the master's conduct.

2. Due to the natural many-sided quality of organic life forms, specialists are regularly faced with discovering non wanted outcomes. Obscure properties could be the reason for these outcomes.

3. As new information and novel idea writes are produced each day in atomic science explore, it is fundamental to apply tech-niques ready to fit this quick developing nature.

4. Machine learning can manage the wealth of absent and loud information from numerous organic situations.

5. Machine learning can manage the gigantic volumes of information produced by novel high-throughput gadgets, with a specific end goal to extricate shrouded connections that exist and that are not perceptible to specialists.

6. In a few organic situations, specialists can just determine input– yield information sets, and they are not ready to depict the general connections between the diverse highlights that could serve to additionally portray how they are related internally. Machine learning can change its inward structure to the current information, creating estimated functions and outcome.

Machine learning algorithms are classified as follows:

- Supervised learning Algorithms: Here a sample dataset called as training set is available which has some input data and the expected output for that data, normally called as classes, the objective here is to build a function to correctly predict output of some unknown input data. Regression is the process of deciding when the desired output is a continuous-value variable. But when the output is a limited set of isolated values, it is called as classification.
- Unsupervised learning Algorithms or clustering: Here the training dataset is divided into various clusters such that the input features which are more identical or related will go in one cluster.
- Semisupervised learning: Here it is combination of unlabeled and labeled data. Here the objective is build a function to correctly predict the output for unknown inputs or the inputs for whom the output is not known. The database here consists labeled data in less quantity and unlabeled data in more quantity.

Apart from supervised and unsupervised learning there are few other learning algorithms like reinforcement learning etc. But as the two supervised and unsupervised are being widely in most of the real world applications in all fields like computational biology, pattern recognition, etc.

II. SUPERVISED CLASSIFICATION

One of the most widely used techniques in Machine learning is supervised classification. It is also called as the approach to predict class. Training dataset is the backbone of this method. Which is made up of a set of N sample data points which are independent of each other viz. $D_N = \{(\mathbf{x}^1, c^1), \ldots, (\mathbf{x}^N, c^N)\}$ Which are taken from a joint, not known probability distribution $p(\mathbf{x}, c)$. Every data point $(\mathbf{x}^i, c^i)$ is characterized by a set of d features which need to be predicted $\{X_1, \ldots, X_d\}$ and variable belonging to class of interest C, which "supervises" the entire procedure which is being carried out.
The dataset is normally classified as training and testing dataset, usually in the ration 70:30. i.e. 70% of the dataset will be used for training purpose and remaining 30% for testing purpose. Once this division is done we train the algorithm or classifier in the supervised mode. i.e. we tell to the classifier algorithm that if this particular input is given what output to produce. This procedure is called as training. And then we test the performance of the classifier with the testing dataset, where we know the class of the input data and we check what prediction is being made by the classifier. Then we compute the efficiency of the classifier i.e how many correct prediction made by the classifier.

Here we quickly describe the main features of famous classifiers like Decision trees also called as classification trees, Naïve Bayes classifier, Nearest neighbor classifier, Support Vector Machines and Artificial Neural Network Classifier.

2.1 Decision Tree Classifier

Decision tree classifiers are one of the very widely used classifiers because of many reasons like they are very simple, fast, effective and have very informative graphical representation. We apply recursive top down process to build the decision tree model which is very easy to understand and check. The decision tree has the top node called as root, and the other nodes are called as internal nodes. The tree is build recursively from root by taking into account one feature at a time, i.e. every node is corresponding to one input parameter. Then we divide the sample by asking recursive questions recursively. The leaf node is the prediction node.

2.2 Naïve Bayes Classifier

Naive Bayes classifier does the classification based on the parameters which are not dependent on each other. The Naïve Bayes classifier can be best described by equation 1.

$$P(C1|P1,P2)=P(P1|C1)P(P2|C2).P(C1)/P(P1)P(P2) \qquad (1)$$

Equation (1) gives the chance of input to belong to class C1 with parameter P1 and P2.
It defines the chance of getting class C1 with parameters P1 and P2 is the fraction whose numerator part is the product of chance of occurrence of P1 with class C1, chance of occurrence of parameter P2 with class C2 and chance of class C1 divided by the product of chance of occurrence of parameter P1 and chance of occurrence of parameter P2. Thus we can see that it is based on the Bayes formula.

2.3 Support Vector Machines

They are a standout amongst the most mainstream grouping procedures being used today. Its vigorous numerical premise and the great correctnesses that it exhibits in numerous genuine errands have set it among specialists' top choices. SVMs outline tests into a higher-dimensional space where a maximal isolating hyperplane among the cases of various classes is built. The strategy works by developing another two parallel hyper-planes on each side of this hyperplane. The SVM strategy tries to discover the isolating hyperplane that augments the territory of detachment between the two parallel hyperplanes. It is expected that a bigger partition between these parallel hyperplanes will infer a superior prescient precision of the classifier. As the amplest territory of division seems to be, indeed, controlled by a couple of tests that are near both parallel hyperplanes, these examples are called bolster vectors. They are likewise the most troublesome specimens to be accurately characterized. As much of the time, it isn't conceivable to consummately isolate all the preparation purposes of various classes; the allowed remove between these misclassified focuses and the most distant side of the partition zone is constrained. In spite of the fact that SVM classifiers are mainstream because of the striking exactness levels accomplished in numerous bioinformatics issues, they are additionally censured for the absence of expressiveness and comprehensi-bility of their scientific ideas.

III. UNSUPERVISED CLASSIFICATION OR CLUSTERING

Unsupervised classification – or clustering – is a key topic in the machine learning discipline. Its beginning stage is a prepare ing database shaped by an arrangement of N autonomous examples DN = (x1, . . ., xN) drawn from a joint and obscure likelihood distri-bution p(x, c). Each specimen is described by a gathering of d pre-dictive factors or highlights {X1, . . ., Xd} and C is a shrouded vari-capable that speaks to the group enrollment of each occasion. Rather than managed order, there is no name that means the class enrollment of an example, and no data is profit capable about the explanation of the database tests in the examination. Bunching, which is likewise casually known as "class revelation," is connected when there is no class to be anticipated, yet rather when the occurrences are to be separated into characteristic gatherings. Once the proper preprocessing steps are performed over the accessible information, grouping strategies parcel the arrangement of tests into sub-sets as per the distinctions/likenesses between them. The distinctive articles are sorted out/taxonomized into gatherings to such an extent that the level of relationship between two items is maximal in the event that they have a place with a similar gathering and negligible generally. Clus-tering mirrors an endeavor to find the fundamental instrument from which examples started.

A key idea in bunching is the sort of separation measure that decides the closeness degree between tests. This will significantly impact the shape and design of the actuated groups, and its race ought to be precisely examined. Normal dis-tance capacities are the Euclidean, Manhattan, Chebychev, or Mahalanobis.

The approval of a grouping structure, both from factual and natural perspectives, is an essential assignment. Factual valida-tion can be performed by evaluating the bunch intelligibility or by checking the heartiness against the expansion of clamor. A natural measure to be considered by any bunching calculation is the minimization of dissimilarities of tests having a place with a similar group, together with the boost of the dissimilarities between the examples of various groups. All things considered, the issue of natural group approval is a profoundly requested undertaking by bio-specialists that still remains an open test. Since a typical normal for natural frameworks is the way that they are not totally portrayed, the race of the best bunch design is viewed as a troublesome undertaking for researcher. Be that as it may, there are cases of late procedures thought to approve bunching structures in various bioinformatics situations.

In numerous bio-situations, accessible specimens are not clarified, which has driven bunching to have been comprehensively used to take care of various bioinformatics issues, for example, gathering homologous arrangements into quality families, joining crests that emerge from a similar peptide or protein in mass spectra examinations, or gathering comparative quality articulation profiles in DNA microarray tests.

Bunching methods assume a focal part in a few bioinformatics issues, particularly in the bunching of qualities in light of their demeanor profiles in an arrangement of hybridizations. In light of the suspicion that expressional similitude suggests some sort of relationship, grouping strategies have opened a path for the examination and explanation of arrangements. As a characteristic augmentation to grouping, the as of late revived biclustering subject has turned into a promising examination territory in bioinformatics. As it is realized that not every one of the qualities of a particular bunch must be gathered into similar conditions, it appears to be normal to expect that few qualities can just change their demeanor levels inside a predetermined subset of conditions. This reality has roused the advancement of particular biclustering calculations for quality articulation information.

3.1 *Partitional Clustering*

Clustering algorithms that have a place with this family relegate every sam ple to a one of a kind group, along these lines giving a segment of the arrangement of focuses. With a specific end goal to apply a partitional clustering calculation, the client needs to settle ahead of time the quantity of bunches in the parcel. Despite the fact that there are a few heuristic techniques for supporting the choice on the quantity of bunches, this issue still stays open. The k-implies calculation is the prototypical and best-known partitional clustering strategy. Its goal is to segment the arrangement of tests into K bunches so that the inside gathering entirety of squares is limited. In its essential frame, the calculation depends on the change country of two natural and quick advances. Prior to the cycle of these two stages begins, an irregular task of tests to K introductory clus-ters is performed. In the initial step, the examples are appointed to groups, generally to the bunch whose centroid is the nearest by the Euclidean separation. In the second step, new group cen-troids are recalculated. The emphasis of the two stages is ended when no development of a protest an alternate gathering will lessen the inside gathering total of squares. The writing gives a high jumper sity of varieties of the K-implies calculation, particularly centered around enhancing the processing times. Its primary downside is that it doesn't restore similar outcomes in two unique runs, since the last setup of bunches relies upon the underlying irregular relegate ments of focuses to K introductory groups.

In fluffy and probabilistic clustering, the examples are not compelled to have a place totally with one bunch. By means of these methodologies, each point has a level of having a place with each of the clus-ters. Guided by the minimization of intracluster difference, the writing demonstrates intriguing fluffy and probabilistic clustering techniques, and the field is as yet open for advance distribution openings.

3.2 Hierarchical Clustering

This is the most extensively utilized clustering worldview in bioinformatics. The yield of a various leveled clustering calculation is a settled and progressive arrangement of allotments/groups spoke to by a tree outline or dendrogram, with singular specimens toward one side (base) and a solitary bunch containing each component at the other (top). Agglomerative algorithms start at the base of the tree, though disruptive algorithms start at the best. Agglomerative techniques construct the dendrogram from the individual examples by iteratively combining sets of groups. Troublesome strategies seldom are connected because of their wastefulness. In view of the straightforwardness and high instinctive level of the dendrogram, the master can deliver a segment into a coveted number of disjoint gatherings by cutting the dendrogram at a given level. This ability to choose the quantity of definite groups to be contemplated has promoted the utilization of various leveled clustering among bio-specialists.
A difference grid with the separation between sets of bunches is utilized to manage each progression of the agglomerative combining process. An assortment of separation measures between bunches is accessible in the writing. The most well-known measures are single-linkage (the separation between two gatherings is the separation between their nearest individuals), finish linkage (characterized as the separation between the two most remote focuses), Ward_s progressive clustering strategy (at each phase of the calculation, the two gatherings that create the littlest increment in the aggregate inside gathering entirety of squares are amalgamated), centroid remove (characterized as the separation between the bunch means or centroids), middle separation (separate between the medians of the groups), and gathering normal linkage (normal of the dissimilarities between all sets of people, one from each gathering).

IV. CONCLUSION

Machine learning algorithms can be effectively used to solve complex biological problems. Those algorithms can be used for the problems like classification of proteins which is one of the demanding problems in computational biology and bioinformatics.

REFERENCES

[1] Prompramote S, Chen Y, Chen Y-PP. (2005) Machine learning in bioinformatics. In Bioinformatics Technologies (Chen Y-PP., ed.), Springer, Heidelberg, Germany, pp. 117–153.
[2] Alpaydin E. (2004) Introduction to Machine Learning, MIT Press, Cambridge, MA.
[3] Mitchell T. (1997) Machine Learning, McGraw Hill, New York.
[4] Parmigiani G, Garett ES, Irizarry RA, Zeger SL. (2003) The Analysis of Gene Expression Data, Springer-Verlag, New York.
[5] Witten IH, Frank E. (2005) Data Mining. Practical Machine Learning Tools and Techniques (2nd ed.), Morgan Kaufmann, San Francisco.
[6] Gentleman R, Carey VJ, Huber W, Irizarry RA, Dudoit S. (2005) Bioinformatics and Computational Biology Solutions Using R and Bioconductor, Springer, New York.