Comparative Analysis of Machine Learning Algorithms on Different Datasets

Conference Paper · April 2019

CITATIONS
18

READS
8,457

4 authors:

Kapil Sethi
BAHRA University
8 PUBLICATIONS **71** CITATIONS
SEE PROFILE

Ankit Gupta
Madeira Interactive Technologies Institute
29 PUBLICATIONS **167** CITATIONS
SEE PROFILE

Gaurav Gupta
Shoolini University
61 PUBLICATIONS **537** CITATIONS
SEE PROFILE

Varun Jaiswal
Shoolini University
97 PUBLICATIONS **937** CITATIONS
SEE PROFILE

Conference Article

Circulation in Computer Science
International Conference on Innovations in Computing (ICIC 2017), pp:87-91

Comparative Analysis of Machine Learning Algorithms on Different Datasets

Kapil Sethi
School of Electrical and Computer Science
Shoolini University
Solan, India

Ankit Gupta
School of Electrical and Computer Science
Shoolini University
Solan, India

Gaurav Gupta
School of Electrical and Computer Science
Shoolini University
Solan, India

Varun Jaiswal
School of Electrical and Computer Science
Shoolini University
Solan, India

ABSTRACT
Machine learning calculations can make sense of how to perform imperative errands by summing up from illustrations. This research aims at comparing different algorithms used in machine learning. Machine Learning can be both experience and explanation-based learning. In this study most popular algorithms were used like Neural Network (NN), k-Nearest Neighbour (KNN) and Support Vector Machine SVM and two datasets were used to check the efficiency of algorithms. Comparative analysis of the classifiers shows that SVM outperforms the other methods with a high accuracy.

Keywords
Machine learning, algorithms, datasets

1. INTRODUCTION
Machine learning (ML) is categorized under artificial intelligence of (AI) which facilitates the computer with efficiency to perform and learn even after not being particularly programmed. ML is a strategy for information examination that robotizes logical model building [1-3]. A present report from the McKinsey Global Institute proclaims that machine learning (a.k.a. data mining or farsighted examination) will be the driver of the accompanying gigantic surge of headway.ML only concentrates on developing computer programs flexible to change whenever expose to new data. Different ML algorithms involve huge potential to be successfully applied in different fields like medicals [1-5], corporates, education, robotics, games and much more [6]. In ML one of the important factors is to make machines able to learn efficiently and effectively [7]. There are an extensive variety of computations which help in making gadgets and strategy in ML. Sometimes these methods create confusion in their applicability in suitable methods and which algorithms gives more accuracy can't know. Researchers have used different algorithms in ML according to expertise, availability and the dataset [8]. Although ML is a discreetly young ground of research [9]. Selecting algorithm in ML for the given datasets (problem) can be tricky. In the ubiquity of machine learning, current relative investigation has been formulated among three existing classifiers (NN, SVM, KNN) utilizing vast datasets. Preprocessing strategies have been utilized to get highlights from bigger informational indexes to prepare the current classifier systems [10]. A comparative analysis is put together in investigating the improved accuracy of classifiers [11]. The objective of these exercises isn't to include new usefulness, however to show the best strategy in examination with these SVM, NN, KNN strategies. Cross validation is used with 90% for Train and 10% for Test the dataset. Also find the accuracy, specificity and sensitivity for comparing ML algorithms.

2. THE USED CLASSIFIERS

2.1 Neural Network (NN)
In the field of information technology (IT), an NN is a structure of hardware and software patterned after the process of neurons in the human mind. NN also functioned with ANN and it is a variability of deep learning skills. These capacities for the most part concentrate on settling complex flag preparing or design acknowledgment issues [12]. Examples are handwriting and voice recognition, data analysis, weather forecast and facial recognition. An NN normally contains a huge number of processors functioning similarly and settled in tiers [13]. The principal level takes the crude contribution of information closely resembling optic nerves in humanoid visual preparing. Each succeeding level gets the yield after the level going before it, somewhat than from the crude contribution to a similar way neurons encourage from the optic nerve get signals from those nearer to it. The last level creates the yield of the framework. Each processing node has its particular minor sphere of information, containing what it has understood and any instructions it was initially planned with or established for itself [14]. The levels exist very interconnected, which implies every hub in level n will be identified with numerous hubs in level n-1 its sources of info and in level n+1, which conveys contribution for those hubs. It might be single or a few hubs in the yield layer, from which the reaction it produces can be perused. NN are prominent for being adaptive [15], which means they modify themselves as they learn from initial training and successive runs provide more info about the world [15]. The best simple learning model is focused on allowance the input streams, which is how each node masses the status of feedback from each of its predecessors. Inputs that contribute to getting right reactions are weighted higher.

Generally, an NN is primarily trained or fed huge bulks of data. It contains giving info and telling the system what the yield ought to be. For example, to form a structure to recognize the expressions of artists, initial training might be a sequence of pictures of artists, non- artists, masks, statuary, animal looks and so on. Individually contribution is attended by the identical identification, like artists' names, "not artists" or "not human" information. Giving the appropriate responses enables the model to modify its inward weightings to procure how to carry out its activity better. NN are sporadically defined in terms of their depth, including how many layers they have between input and output, or the model's so-called hidden layers. Disparities on the standard NN design allow various forms of forwarding and backward propagation of information among tiers. NN was first created as a share of

the wider research effort around AI, and they continue to be important in that space, as well as in research around humanoid perception and realization.

For accumulating the NN we use this formula 1.

$$a_j^l = \sigma\left(\sum_k w^l{}_{jk}\, a_k^{l-1} + b_j^l\right) \quad (1)$$

In this formula $w^l{}_{ik}$ to denote the mass for the linking from the k^{th} neuron in the $(l-1)^{th}$ layer to the j^{th} neuron in the l^{th} layer. Explicitly, we use b^{lj} for the bias of the j^{th} neuron in the l^{th} layer. And we use a^{lj} for the start of the j^{th} neuron in the l^{th} layer.

2.2 Support vector machines (SVM)

In ML, SVM are supervised learning models associated with learning algorithms that inspect data used for classification and regression analysis [16]. Determined a settled of activity cases, each show as going to one or the new of two gatherings, in SVM preparing calculation develops a model that apportions new cases to one gathering or the other, making it a non-probabilistic parallel direct classifier [9]. When facts are not categorized, supervised learning is not possible, and an unsupervised learning approach is compulsory [17], which efforts to invention normal clustering of the data to groups, and then map new data to these formed groups. The clustering algorithm which delivers an enhancement to the SVM is called support vector clustering (SVC) and is often used in trade applications either when facts are not categorized or when only some facts are categorized as a preprocessing for a classification pass [18]. The mechanism of classifying the data into different classes by definition a line which splits the training files into classes. There are a few straight hyperplanes, SVM calculation tries to augment the separation in the focal of the few classes that are mind boggling and this is said as edge augmentation [19]. If the line makes the most of the space among the classes is recognized, the probability to simplify well to unobserved data is increased. There are two categories in SVM:

2.2.1 Linear SVM (L-SVM)

In L-SVM the training statistics i.e. classifiers are separated by a hyperplane.

$$\frac{1}{m}\sum_{i=1}^{m} l\ (w\,.\,x_i + b\,.\,y_i) + \| w \|2 \quad (2)$$

2.2.2 Non-Linear SVM (NL-SVM):

In NL-SVM it is not thinkable to discrete the training statistics with a hyperplane. For example, the training statistics for Face recognition consists of a group of images that are faces and another group of images that are not faces (in other words all other images in the world except faces). Under such conditions, the preparation measurements are excessively perplexing and troublesome, making it impossible to find a delineation for each component vector [10]. Isolating the typical of countenances directly after the arrangement of non-confront is a many-sided assignment.

2.3 K- nearest neighbors (KNN)

KNN is a modest algorithm that stores all accessible suitcases and classifies new suitcases based on a similarity measure. KNN has symmetrical names (a) Memory-Based Reasoning [20] (b) Example-Based Reasoning (c) Instance-Based Learning (d) Case-Based Reasoning and (e) Lazy Learning.

KNN utilized for relapse and grouping for prescient issues [10]. Be that as it may, it is broadly utilized as a part of grouping troubles in the business. To assess any procedure, we by and large take a gander at 3 critical angles:

1. Ease to interpret output
2. Calculation time
3. Predictive Power

3. EXPERIMENTAL DATASET AND METHODOLOGY

In the present study two different datasets are used. First data is used quality of red and white wine with 6500 rows and 13 columns [21]. The second one is biodegradable chemical with1055 rows and 13 columns [22]. All information procured from UCI machine learning storehouse.

The methodology for the classification of these datasets is displayed in Figure 1. The analysis has been performed in MATLAB platform running on (Intel i5 processor) with 3 GB RAM installed. The different data sets are taken as input for feature extractor and classification algorithm. The datasets are passed through a sequence of pre-processing blocks.

3.1 Different data

In the current study combination of two datasets have been used. These two datasets were downloaded from ML repository UCI.

3.2 Pre-processing

After the gathering of data next phase is to perform the preprocessing on the collected data. It is the technique that changes the raw data to the understandable format.

3.3 Filtering

After preprocessing of data next phase is to filter the data.

3.4 Normalization

It is a stage in which all the values are changed to values between 0 and 1. The reason for standardization is to attract the information to an alternate scale.

3.5 ML algorithms

After normalization use different algorithms used such as NN, SVM and KNN.

3.6 Validation

In the validation10-fold cross-validation method is used.
To create the training and testing sets, all normalized features are randomized before they can be used to train the classifier networks; *K*NN, NN, and SVM. *K*NN is evaluated for three nearest neighbors (*k*=3) with distance metric as Euclidean. The separation is computed between test information and every case of preparing information. This separation

controlled by various highlights utilized for the grouping. These distances along with the supervised classes are sorted in an ascending order, then depending upon the value of *k*, the *k* NN is taken out and maximum class allotted to those samples is selected as the class of the test data. ML-based characterization is performed by processing the odds of the test occasions and is allotted a class for which the likelihood is most astounding.

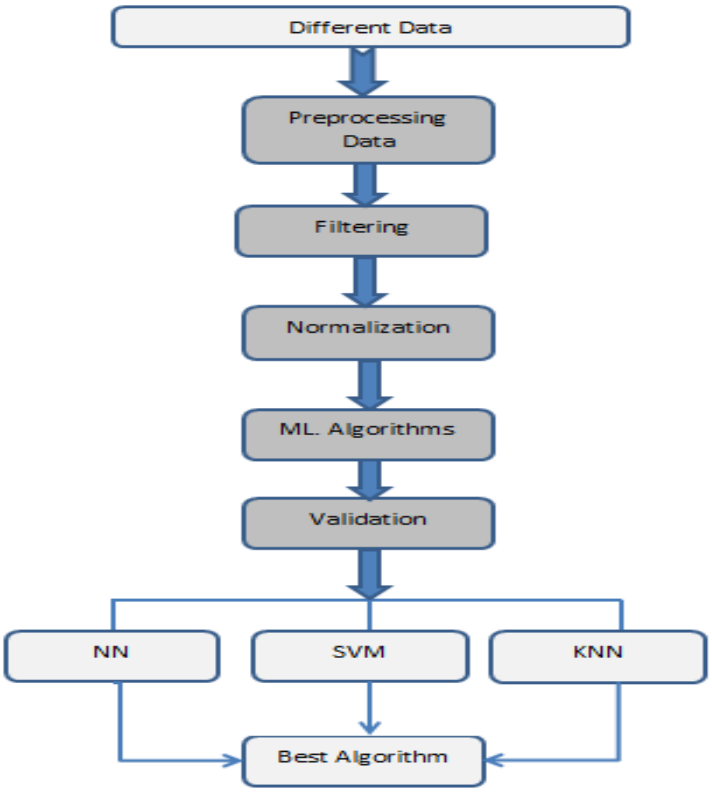

Fig 1: Schema of experimentation methodology

4. RESULT AND DISCUSSION

In current work two different datasets (wine and biodegradable chemical) were used. To approve and assess the execution of classifiers, 10-cross validation strategy is utilized. The accuracy of each fold is computed and the classifiers models with the highest accuracy are selected for classification. All accuracy of the dataset is presented in Table 1 and Roc curves shows in Figure 2. According to the Table 1 SVM method was used where biodeg dataset and wine dataset was applied on 90% train and 10% test. We got different amazing results with accuracy 88.57%, specificity is 0.69 and sensitivity is 0.99 in biodeg dataset and accuracy 99.38%, specificity is 0.97 and sensitivity is 0.98 in wine dataset. Similarly, Neural Network method was used where results were observed for biodeg dataset and wine dataset on train 90% and test 10%. The accuracy is 87.70%, specificity is 0.77 and sensitivity is 0.97 in biodeg dataset and accuracy is 99.00%, specificity is 0.78 and sensitivity is 0.99 in wine dataset. For the nearest neighbour method, biodeg and wine dataset were used where the accuracy for the biodeg is 83.81%, 0.79 specificity and 0.84 sensitivity. In the wine dataset the accuracy is 95.23%, 0.15 specificity and 0.85 sensitivity. The Table 1 indicates SVM gives the best outcomes as contrast with the others and it gives the higher precision with 99.38.

In the past investigation of wine datasets three diverse grouping calculations SVM, NN and MR were connected with 5-overlay cross-approval technique. It was discovered that SVM exactness is 87.9% and two times superior to the others [21]. In the other dataset of bio degradable additionally utilized three calculations KNN, PLSDA and SVM to discover blunder, specificity and affect-ability. 5- Cross validation technique was utilized and it was discovered that KNN has the most minimal blunder then the other two that is 0.17, affect-ability is 0.75 and specificity are 0.91[22]. Relatively to current work three different classification algorithms KNN, NN and SVM is used. According to the current results SVM classifier has the highest accuracy of 99.38% from the others.

Table 1. Accuracy, specificity, sensitivity of two datasets viz. biodeg (upper value), wine (lower value).

Method (Datasets)	Train 90%	Test 10%	Accuracy	Specificity	Sensitivity
Biodeg SVM	950	105	88.57	0.69	0.99
Wine	5850	650	99.38	0.97	0.98
Biodeg Neural Network	950	105	87.70	0.77	0.97
Wine	5850	650	99.00	0.78	0.99

Biodeg Nearest Neighbour	**950**	**105**	**83.81**	**0.79**	**0.84**
Wine	**5850**	**650**	**95.23**	**0.15**	**0.85**

According to Figure 2 the A, b, C, D, E and F is a ROC curve (Receiver operating characteristic) of biodeg and wine dataset of applied methods that is SVM, NN and KNN. The ROC permits to make bends and a total affectability/specificity report. The ROC bend is a major device for indicative test assessment. In the above figures X-axis lies on 0 to 1 and it shows the False positive rate (sensitivity) and Y-axis lies on 0 to 1 and it shows the True positive rate (specificity). Late years have seen an expansion in the utilization of ROC diagrams in the machine learning group, due to some degree to the acknowledgment that basic characterization exactness is regularly a poor metric for measuring execution [23]. Figure (B) NN_wine is the perfect Classifier ROC curve as compare the others.

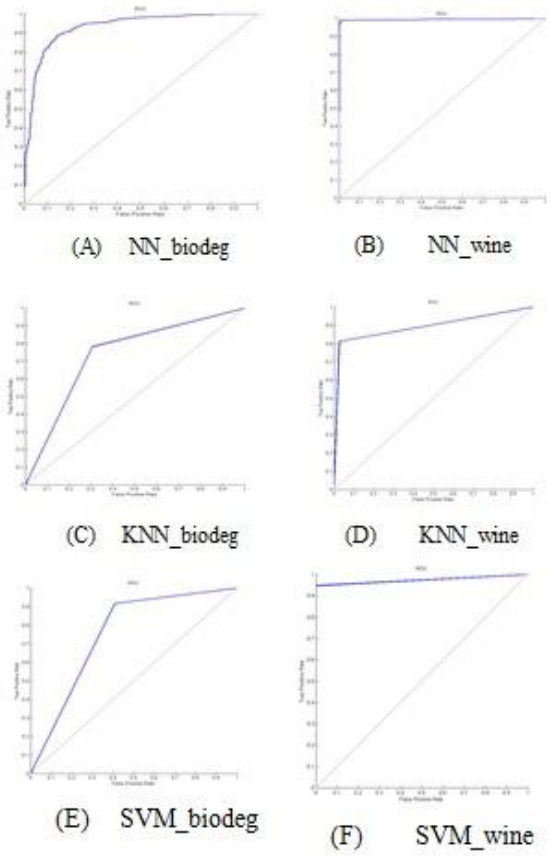

Fig 2. Receiver operating characteristic curves A, B, C, D, E and F.

The formula is to calculate the sensitivity and specificity is given below.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FN}}$$

Where TP, FP, TN, and FN symbolize the number of True positives, False positives, True negatives, and False negatives, respectively.

5. CONCLUSION

As per current examination the relative investigation of KNN, SVM, and NN has been executed and the execution metric mirrors the execution of SVM better when contrasted with the other broke down classifier. The outcomes certainly demonstrate the origination that the highlights required for the preparation of the model for classifier ought to be strong and unmistakable with the goal that different procedures of unsupervised and directed learning can be investigate to improve the execution. The accuracy estimation of SVM strategy was observed to be close to 99.38% it shows the higher effectiveness of SVM for forecast of models or apparatuses for future forthcoming. The SVM based models and instruments might be useful in the field of restorative sciences, law, fund, governmental issues, medication, instruction, and different fields.

6. REFERENCES

[1] Sharma, L., Gupta, G. and Jaiswal, V., 2016, December. Classification and development of tool for heart diseases (MRI images) using machine learning. In *Parallel, Distributed and Grid Computing (PDGC), 2016 Fourth International Conference on* (pp. 219-224). IEEE.

[2] Chauhan, D. and Jaiswal, V., 2016, October. An efficient data mining classification approach for detecting lung cancer disease. In *Communication and Electronics Systems (ICCES), International Conference on* (pp. 1-8). IEEE.

[3] Negi, A. and Jaiswal, V., 2016, December. A first attempt to develop a diabetes prediction method based on different global datasets. In *Parallel, Distributed and Grid Computing (PDGC), 2016 Fourth International Conference on* (pp. 237-241). IEEE.

[4] Pal, T., Jaiswal, V. and Chauhan, R.S., 2016. DRPPP: A machine learning based tool for prediction of disease resistance proteins in plants. *Computers in biology and medicine*, *78*, pp.42-48.

[5] Jaiswal, V., et al., Jenner-predict server: prediction of protein vaccine candidates (PVCs) in bacteria based on host-pathogen interactions. BMC bioinformatics, 2013. **14**(1): p. 211.

[6] Jaiswal, V., Chanumolu, S.K., Gupta, A., Chauhan, R.S. and Rout, C., 2013. Jenner-predict server: prediction of protein vaccine candidates (PVCs) in bacteria based on host-pathogen interactions. *BMC bioinformatics*, *14*(1), p.211.

[7] Das, S., Dey, A., Pal, A. and Roy, N., 2015. Applications of Artificial Intelligence in Machine Learning: Review and Prospect. *International Journal of Computer Applications*, *115*(9).

[8] Sebastiani, F., 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1), pp.1-47.

[9] Cunningham, S.J., Littin, J. and Witten, I.H., 1997. Applications of machine learning in information retrieval.

[10] Mitchell, T.M., 2006. *The discipline of machine learning* (Vol. 3). Carnegie Mellon University, School of Computer Science, Machine Learning Department.

[11] Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.F. and Dennison, D., 2015. Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems* (pp. 2503-2511).

[12] Portugal, I., Alencar, P. and Cowan, D., 2015. The use of machine learning algorithms in recommender systems: a systematic review. *arXiv preprint arXiv:1511.05263*.

[13] Cost, S. and Salzberg, S., 1993. A weighted nearest neighbor algorithm for learning with symbolic features. *Machine learning*, *10*(1), pp.57-78.

[14] Del Pezzo, E., Esposito, A., Giudicepietro, F., Marinaro, M., Martini, M. and Scarpetta, S., 2003. Discrimination of earthquakes and underwater explosions using neural networks. *Bulletin of the Seismological Society of America*, *93*(1), pp.215-223.

[15] Shalev-Shwartz, S. and Ben-David, S., 2014. *Understanding machine learning: From theory to algorithms*. Cambridge university press.

[16] Wagstaff, K., 2012. Machine learning that matters. *arXiv preprint arXiv:1206.4656*.

[17] Bennett, K.P. and Parrado-Hernández, E., 2006. The interplay of optimization and machine learning research. *Journal of Machine Learning Research*, *7*(Jul), pp.1265-1281.

[18] Caruana, R. and Niculescu-Mizil, A., 2006, June. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning* (pp. 161-168). ACM.

[19] Javidi, B., 2002. *Image recognition and classification: algorithms, systems, and applications*. CRC Press.

[20] Domingos, P., 2012. A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), pp.78-87.

[21] Cortez, P., Cerdeira, A., Almeida, F., Matos, T. and Reis, J., 2009. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, *47*(4), pp.547-553.

[22] Mansouri, K., Ringsted, T., Ballabio, D., Todeschini, R. and Consonni, V., 2013. Quantitative structure–activity relationship models for ready biodegradability of chemicals. *Journal of chemical information and modeling*, *53*(4), pp.867-878.

[23] Fawcett, T., 2006. An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), pp.861-874.