IJCSE International Journal of Computer Sciences and Engineering Open Access

Survey Paper Vol.-7, Issue-4, April 2019 E-ISSN: 2347-2693

"Breast Cancer Diagnosis and Classification Using Support vector machines With Diverse Datasets"

Vikas S[1*], Thimmaraju S N[2]

[1,2]Dept. of MCA, VTU PG Centre Mysuru, Karnataka, India

Corresponding Author: vikas.smg@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i4.442446 | Available online at: www.ijcseonline.org

Accepted: 09/Apr/2019, Published: 30/Apr/2019

Abstract- Breast cancer is the most prevalent cancer among women around the world. However, increased survival is due to the dramatic advances in the screening methods, early diagnosis, and breakthroughs in treatments. Different strategies of breast cancer classification and staging have evolved over the years. Intrinsic (molecular) sub composing is fundamental in clinical preliminaries and well comprehension of the sickness of the disease.

To analyze machine learning systems have been utilized to define a set trained with the "bagging" method. Support vector machines (SVM) have been appeared to outflank numerous related methods. However, there have been very few studies focused on examining the classification performances of different classification.

The trial comes about demonstrate that SVM classifier can be the better decision for classification, where accuracy of the algorithm is improved by tuning the parameters of the dataset.

Keywords-Machine Learning, Support Vector Machine(SVM).

I. Introduction

Many big data use cases have been realized, which create additional value for companies, end users and third parties. Currently, real time data is gathered from millions of end users via popular social networking services.

For example, Twitter uses collected data for real time query suggestion and spelling corrections of their search algorithm. Analysis of collected data also increases understanding of consumers, which is an important asset for the big data companies. Value from data can also be extracted with other applications.

SVM has shown itsbunches of exceptional capacity, particularly in characterization issues and especially in classification problems. Its basic design philosophy is to maximize the classification boundaries and its basic purpose is to maximize the hyper-plane.

To reduce the support vector machine time and space complexity, many improved calculation has been connected effectively and algorithm has been applied successfully. One method is to obtain low-order approximation of the nuclear matrix by greedy algorithm [1], or sample [2], or decomposition matrix. If dimension decomposed nuclear matrix is still very high, resulting in SVM training efficiency is still very low. Another method is to improve the efficiency of SVM algorithm block.

We have explored different avenues regarding various parameters related with the utilization of the SVM algorithm that can impact the results. These parameters include choice of kernel functions, the standard deviation of the Gaussian kernel, relative weights associated with slack variables to account for the non-uniform conveyance of marked information, and the quantity of preparing precedents.

The data used in this experiment was obtained from the UCI machine learning repository [11] and described by Dr. William H. Wolberg. Information have been utilized in some examination. [14] We discussed the effect of 31 characteristic parameters on the condition of breast cancer and the influence of the involved parameter on the performance of the SVM models.

We visualize the data using density plots to get a feeling of the information conveyance. in the meantime, the comparison between the performance of SVMs and other techniques was performed using these data. The problem is to predict the state of breast cancer. In this database, there are 569 pieces of samples, and every sample is communicated by 31 characteristic parameters.

II. Overview Of Our Approach

We shall consider SVMs in the binary classification setting. We are given training data {x1 ... xn} that are vectors in some space X ⊆ Rd. We are also given their labels {y1 ...yn} where yi ∈ {−1, 1}. In their simplest form, SVMs are hyper-planes that separate the training data by a maximal margin (see Fig. 1a) . All vectors lying on one side of the hyper-plane are labeled as −1, and all vectors lying on the other side are labeled as 1. The training instances that lie closest to the hyper-plane are called support vectors. More generally, SVMs allow one to project the original training data in space X to a higher dimensional feature space F via a Mercer kernel operator K. In other words, we consider the set of classifiers of the form:

$f(x) = ({}^{n}\sum_{i=0} (\alpha iK(xi, x))$

When K satisfies Mercer's condition (Burges, 1998) we can write: K(u, v) = Φ(u) · Φ(v) where Φ : X→F and "·" denotes an inner product.
We can then rewrite f as:

$f(x) = w \cdot \Phi(x)$, where $w = {}^{n}\sum_{i=0} \alpha i\Phi(xi)$. (2)

Thus, by using K we are implicitly projecting the training data into a different (often higher dimensional) feature space F. The SVM then computes the αis that correspond to the maximal margin hyperplane in F. By choosing different kernel functions we can implicitly project the training data from X into spaces F for which hyperplanes in F correspond to more complex decision boundaries in the original space X.

Selection criteria
In a research or production environment, the choice of machine learning packages or specific algorithms will come down to a variety of different factors, for the most part reliant on the requirements of the particular gathering or venture. A number of authors have tackled this area, including [28–29]. Based in part on these studies, we offer a list of important considerations for evaluation of machine learning tools. These are introduced in no specific request, since the prioritization of these components will be reliant on specific use cases.

• *Scalability* This should be considered with regards to both the size and complexity of the data. Scalability should be looked at in both directions, as some of the best tools for big data perform poorly on small data, and vice versa. This is also true for other data characteristics, such as dimensionality.
One should consider what their data looks like now, as well as what data they might be working with in the future, so as to decide whether a specific toolbox will be proper.

•*Speed the* biggest factor affecting speed is which processing platform the library or algorithm is running on rather than the library or algorithm itself. Speed may not be important for every project. If models do not require frequent updating, a batch system might be favored for its effortlessness, yet for models that are refreshed frequently, this might be a significant concern.

•*Coverage this* refers to the range of options contained in the toolkit in terms of different classes of machine learning as well as variety of implementations in each class. None of the available tools for big data provide a selection as comprehensive as some non-distributed frameworks such as Weka, but their degree may go from just a couple of calculations to around two dozen. The same numbers of the instruments are hard to set up and learn, it is critical to think about future needs just as current.

Related work
Single-machine frameworks Many machine learning researchers carry out their work on a single—often GPUequipped—computer [19, 20], and many flexible single machine frameworks have emerged to support this scenario. Caffe [16] is a superior system for preparing decisively determined convolutional neural systems that keeps running on multicore CPUs and GPUs.

Batch dataflow systems Starting with MapReduce [22], batch dataflow systems have been applied to a large number of machine learning algorithms [11], and more recent systems have focused on increasing expressivity and performance.

The principal limitation of a batch dataflow system is that it requires the info information to be permanent, and the majority of the sub computations to be deterministic, so the framework can re-execute sub computations when machines in the bunch fall flat.

III. PROPOSED MODEL

The experimental procedure is based on the following Technique.

a. The given dataset is isolated into 85% preparing and 15% testing sets dependent on the 10-overlay cross approval system [28].
b. We visualize the data using density plots to get a sense of the information conveyance.
c. The testing set is encouraged into the built classifiers preceding examination of their classification accuracy, precision, and F-measure rates.
d. The classifier preparing times are likewise contrasted with investigate the computational complexities of preparing distinctive classifiers. A graph is plotted to compare performance of SVM, KNN and Gaussian Naive Bayes.

The Dataset
In this paper, a breast cancer datasets is utilized, which is accessible from the UCI machine learning repository
(available at: http://archive.ics.uci.edu/ml/)

This relatively small scale dataset, which is composed of 659 data samples and each data sample has 31 distinct highlights.

The Classifier Design

There are four single classifiers to be specific, linear SVM, KNN and Gaussian Naive Bayes. In addition, to evaluate the performance of the different SVM classifiers, in addition to the classification accuracy,precision, and the F-measure rate, the time that is spent preparing every classifier is likewise also compared.

Working of the proposed system

The working of the system is depicted as follows:

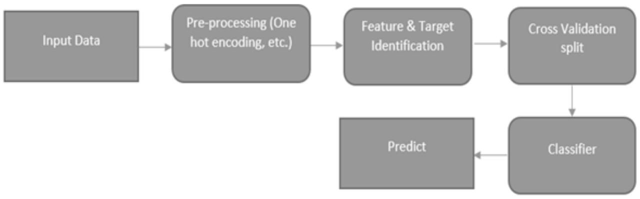
Input Data
Pre-processing (One hot encoding, etc.)
Feature & Target Identification
Cross Validation split
Classifier
Predict

Fig 1.Flowchart of system

IV. EXPERIMENTAL RESULTS

In this section, the results of the classification are reported. To apply our classifiers and assess them, we apply the 10-crease cross approval test which is a method utilized in assessing prescient models that split the first set into a preparation test to prepare the model, and a test set to assess it.

After applying the pre-processing and preparation methods, we try to analyse the data visually and figure out the distribution of values in terms of effectiveness and adequacy.

Density Plots

We can see that perhaps the attributes perimeter, radius, area, concavity; compactness may have an exponential distribution. W We can likewise observe that maybe the texture and smooth and symmetry attributes may have a Gaussian or nearly Gaussian distribution.

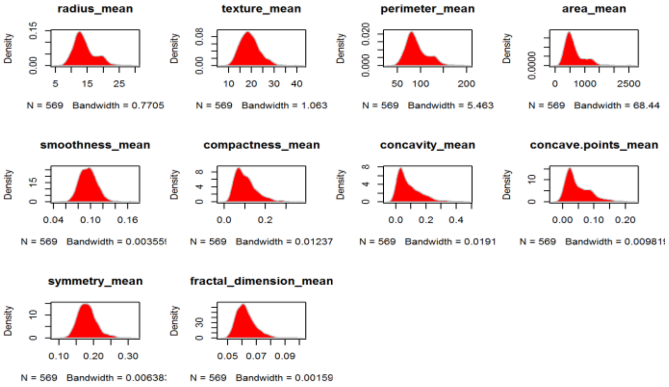

Fig.2

Scatter Plot

A matrix of the visual representation of the relationship between the 6 most highly correlated variables:1.radius_mean2.parameter_mean3.area_mean 4. compactness_mean 5. concavity_mean 6. concave_points_mean.

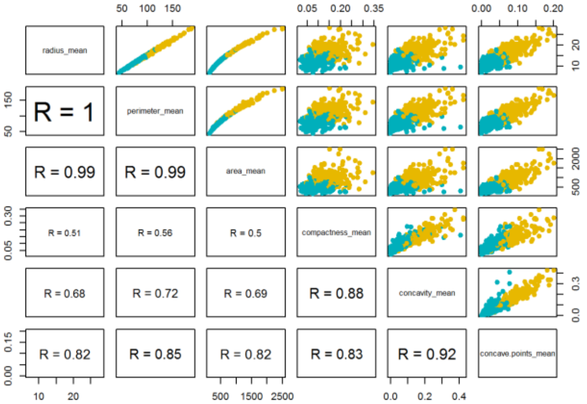

Fig.3

We can clearly see that we can easily distinguish the difference between Malignant and Benign. The majority of the kind perceptions are focused in the left lower quadrant of the diagram while the threatening perceptions are focused in the correct upper quadrant. As well as some variable interactions have an almost linear relationship.

Efficiency

Once the predictive model is built, we can check how productive it is. For that, we compare the accuracy measures based on precision, recall, F1- score values for SVM, Gaussian NB and k-NN. To better understand efficiency, Fig. 3 shows the arrangement report of our classifiers that better represent the exactness of every classifier. It gives a graphical diagram that represents the execution of various classifiers.

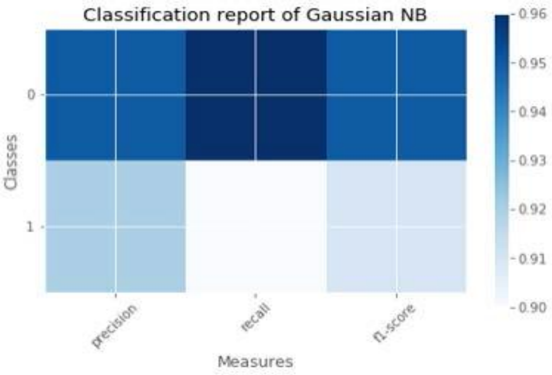

Fig.4

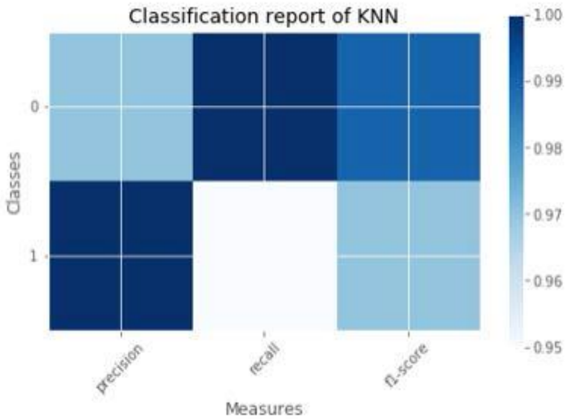

Fig.5

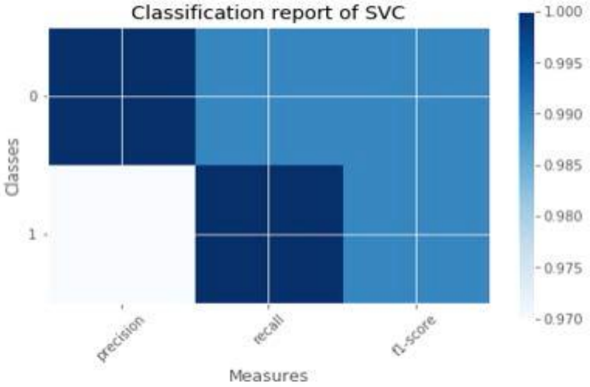

Fig.6

From the above plots we can easily select optimal models and discard others to best classification. Since Confusion matrices represent a useful way for evaluating classifier, each row of Table 3 represents rates in an actual class while each column shows predictions.

V. CONCLUSION

To analyze medicinal information, various data mining and machine learning methods are available. An imperative assignment in the field of machine learning is to build accurate and computationally streamlined classifiers for Medical applications. In this study, we employed four main algorithms: SVM, NB, K-NN on the Wisconsin Breast Cancer datasets.

We tried to compare efficiency and effectiveness of those algorithms in terms of accuracy, precision, recall and F-measures to find the best classification accuracy.

SVM reaches and accuracy of **98.41%** and outperforms, therefore, all other algorithms. And Previous research paper have mentioned about the Accuracy was **97.13%.**

In conclusion, SVM algorithm has proven its efficiency and accuracy in breast cancer diagnosis and has achieved the optimum performance in terms of precision and low error rate.

REFERENCES

[1]. U.S. Cancer Statistics Working Group. United

[2]. States Cancer Statistics: 1999–2008 Incidence and Mortality Web-based

[3]. Report. Atlanta (GA): Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute; 2012.

[4]. Siegel RL, Miller KD, Jemal A. Cancer Statistics , 2016.

[5]. Noble WS. What is a support vector machine? Nat Biotechnol.

[6]. Rish I. An empirical study of the naive Bayes

[7]. classifier. IJCAI Work Empir methods ArtifIntell. 2001.

[8]. Quinlan JR. C4.5: Programs for Machine Learning.

[9]. Larose DT. Discovering Knowledge in Data.

[10]. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2004.

[11]. X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang,H. Motoda, G. J. Mclachlan, A. Ng, B. Liu, P. S. Yu, Z. Z. Michael, S.

[12]. David, and J. H. Dan, Top 10 algorithms in data mining. 2008.

[13]. Datafloq - Top 10 Data Mining Algorithms,

[14]. Demystified.

[15]. 44. Accessed December 29, 2015.V. Chaurasia and S. Pal, "Data Mining Techniques : To Predict and Resolve Breast Cancer Survivability," vol. 3,2014.

[16]. Djebbari, A., Liu, Z., Phan, S., AND Famili, F. International journal of computational biology anddrug design (ijcbdd). 21st Annual Conference on Neural Information Processing Systems (2008).

[17]. S. Aruna and L. V Nandakishore, "KNOWLEDGE BASED ANALYSIS OF VARIOUS STATISTICAL TOOLS IN DETECTING BREAST,"

[18]. A. C. Y, "An Empirical Comparison of Data Mining Classification Methods," vol. 3, no. 2, 2011.

[19]. A. Pradesh, "Analysis of Feature Selection with Classification : Breast Cancer Datasets," Indian J. Comput. Sci. Eng., vol. 2, no.5, 2011.

[20]. Thorsten J. Transductive Inference for Text

[21]. Classification Using Support Vector Machines.

[22]. Icml. 1999.

[23]. L. Ya-qin, W. Cheng, and Z. Lu, "Decision tree based predictive models for breast cancer

[24]. survivability on imbalanced data," 2009.

[25]. D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," Artif. Intell. Med., vol. 34, pp. 113–127, 2005. W. Version, "Machine Learningwith WEKA," 2004.

[26]. "UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set." [Online].

[27]. "SUGI 31 Statistics and Data Analysis Receiver Operating Characteristic (ROC) Curves MithatGönen , Memorial Sloan-Kettering Cancer Center SUGI 31 Statistics and Data Analysis FN + FP,", 2001.

[28]. Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In Proceedings of NAACL-2013, pages 380–390.

[29]. Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu

Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830.

[30]. M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma,M. McCauley, M. Franklin, S. Shenker, and I. Stoica.Resilient Distributed Datasets: A fault-tolerantabstraction for in-memory cluster computing. In NSDI, 2012.

[31]. M. Bostock, V. Ogievetsky, and J. Heer. D3: Data-Driven Documents. In InfoVis, 2011.

[32]. Pandey, Neha, B. K. Singh, and Ankur Singh Bist. "A novel feature learning for image classification using wrapper approach in GA." Signal Processing and Integrated Networks (SPIN), 2015 2nd International Conference on. IEEE, 2015.

[33]. M. Li, T. Zhang, Y. Chen, and A. J. Smola. Efficient mini-batch training for stochastic optimization. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, pages 661–670, New York, NY, USA, 2014.ACM. ww.cs.cmu.edu/˜muli/file/minibatch sgd.pdf.

[34]. C. J. Maddison, A. Huang, I. Sutskever, and D. Silver. Move evaluation in Go using deep convolutional neural networks. arXiv preprint arXiv:1412.6564, 2014. arxiv.org/abs/1412.6564.

[35]. F. McSherry, M. Isard, and D. G. Murray. Scalability! But at what COST? In Proceedings of the 15th USENIX Conference on Hot Topics in Operating Systems, HOTOS'15, Berkeley, CA, USA, 2015. USENIX Association. www.usenix.org/system/files/conference/hotos15/hotos15-paper-mcsherry.pdf.

[36]. P. Moritz, R. Nishihara, I. Stoica, and M. I. Jordan. SparkNet: Training deep networks in Spark. In International Conference on Learning Represntations, 2016. arxiv.org/abs/1511.06051.

[37]. Kriti Jain1, Megha Saxena2andShweta Sharma3 "Breast Cancer Diagnosis Using Machine Learning Techniques" IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 5 Issue 5, May 2018 ISSN (Online) 2348 – 7968

AUTHORS PROFILE

Mr.Vikas S. received M.Phil degree in Computer Science in the year 2009 and Master of computer Applicaions (MCA) in the year 2007 from Visvesvaraya Technological University and Bachelors Degree in Computer Science in the year 2004 from kuvempu University. He is currently working as Assistant Professor in the **Department of MCA, Visvesvaraya Technological University,** PG Center, Mysore, Karnataka, where he is involved in research and teaching activities. He is having 11 years of teaching experience and 02 years of Industrial experience. He is a Life member of India Society for Technical Education (LMISTE), Computer Society of India (CSI) and Doing Research work on the Area **Big data Analytics**.

Dr. Thimmaraju S N, he is presently a professor and heading the **Department of Master of Computer Application, Visvesvaraya Technological University,** PG Center, Mysore, Karnataka, he has received his Ph.D degree from **Visvesvaraya Technological University(VTU)**, Belgaum in the year 2010, M.E., degree in Computer Science and Engineering from University Visvesvaraya College of Engineering (UVCE), Bangalore in 2002 and Bachelors Degree in Computer Science and Engineering from PESCE, Mandya in the year 1999. He is involved in research and teaching activities. His major areas of research are Computer Networks, WSN's and Graph theory. He is having 17 years of teaching experience. He has published around 15 research papers which include International Journals, International Conferences and Notional Conferences.