Journal of Scientific & Industrial Research
Vol. 80, May 2021, pp. 457-462

A Novel Approach for Multi Variant Classification of Medical Data in Short Text

M Supriya Menon* and Pothuraju Rajarajeswari
Dept of CSE, Koneru Lakshmaiah Education Foundation, AP, India

Received 25 November 2020; revised 03 January 2021; accepted 08 April 2021

Data Mining Techniques has attained its momentum in several areas, and its efficient performance in decision support has outperformed and made it a reliable choice. The medical world is one such empirical domain in which a perfect decision at right time would turn out to be a lifesaver. Medical data figures out to be majorly multi-dimensional, where relevant feature extraction is a challenging factor. Several classification approaches like SVM, Decision Trees, and Naive Based are considered to handle these profound challenges. One such challenge discussed in our paper emphasizing on Medical decision support system with Machine Learning (ML) Methodology considering diseases and treatments with their semantic relations in the document of Pub med abstracts. The proposed Multi variant classification framework aims at reducing data into attributes using PCA Transformation infusion with an efficient classification Algorithm - CNB. Our computed results are comparatively successful in attaining ultimate outcomes concerning performance metrics like Accuracy, Precision, Recall, and Time. The strength of our work lies in presenting an efficient approach for elevating enhanced decisions in Health care.

Keywords: Complement Naïve Bayes (CNB) Algorithm, Data mining, Multi variant classification, Principal Component Analysis (PCA)

Introduction

Despite various domains of research study, health care is often a wing of interest. Several web-based tools like Microsoft Health Vault, Google Health, Samsung health, etc. are standing up to the mark in educating people to handle and manage their health status. The medical field is currently relying on evidence-based medicine, which is a framework supporting decision making in healthcare.[1,2] It deduces information based on patient's family history, scientific proof, and clinical results, which are stored as records for future reference. These records are termed as the patients' e-records or Electronic health record lending benefits like building clinical databases, handling medication, extract health-specific treatment options, thereby contributing a faster and reliable information source.[3,4] One such richest and accepted information resource is Pub med.[5]

Pub med, a free resource encouraging the search and access of reviewed biomedical and life science, literature for improved health decisions. Pop outs are challenging and cumbersome as Pub meds holds more than 30 million abstracts. Their work is focused on retrieving more reliable productive and specific conclusions from the medical dataset subjecting it to 3 phases of execution, includes preprocessing, exploring semantic relation between disease treatment with tokens like the cure, prevent, and side effects and applying classification for profound solutions.

Our objective is to provide an enhanced outcome with classification techniques in combination with the machine learning approach supporting transformation, applicable for depicting suitable medical information in a short text. The model is an exemplary in focusing on personalized medicine to meet specific patient needs, resulting in an ultimate treatment study for certain diseases. Therefore, our research aims to review various representation approaches infusion with several learning techniques to analyze and bring out relations in biomedical domain from the dataset.

Classification Algorithms

Classification is an important phase in data mining tasks, can be performed either on structured or unstructured data allowing them to be categorized into respective defined classes. Different types of classification algorithms are available each aiming at their advantage based on the area of application. Any classification algorithm accepts pre-processed input, assigning it to a classifier for mapping onto a specific category. Few types of classification algorithms with their representative models are decision models (decision trees) logistic regression model (logistic regression algorithm), linear classifier like support

*Author for Correspondence
E-mail: supriyamenon05@gmail.com

vector machine (SVM)[6], and probabilistic models (Naïve Bayes and Complement Naïve Bayes).[7] There are many more classifiers available, but emphasizing few as they reflect learning algorithms which promise their work on both long and short text. Few algorithms and their deficiencies in comparison with proposed are discussed below.

Decision Trees[4] are one such simple one ought to understand and visualize results in a sequence of rules for data classification. Logistic regression is an ML-based algorithm modeled with logistic function for independent variables constrained to handle binary data. These algorithms despite highly representative for categorical and numerical attributes suffer from instability problem. A minute change in data reflects a large change in structure of decision tree, leading to inaccuracy. The computed complexity of probabilities of different branches, identifying their best split, and selecting optimal methods for pruning question the expertise of the algorithm.

SVM offers good support in high dimensional space by representing training data in data space as categories and mapping new data to the same space and classifying them based on the side they occupy. It fails to handle, as its performance over large datasets is computationally inefficient. They do not support probability estimates, a parameter of concern in many classification algorithms. At times when the count of features is more than the count of samples, the algorithm is vulnerable to over fitting.

Naïve Based classification algorithm[8] deduced from Bayes theorem with a class independence property, attains high efficiency and speed in contrast with other sophisticated algorithms as they work well with less amount of training data. This family of probabilistic classifier builds Machine learning models at rapid rate for predictions. Yet, it suffers from the fact of skewed bias data and feature independent assumption that are well addressed in CNB using normalization parameters.

Complementing NB generates the CNB algorithm opted for imbalanced data sets and assuring doubtless performance enhancement on the Text Classification task. This factor forces us to make use of CNB in our proposed framework as it relies on extracting more accurate results on short text within text-based data. ML algorithm[1] uses Predictive Models that sound to offer a good hands-on hand in extracting features from text data thereby improving results. These predictive algorithms are well versed in figuring the absolute class labels by using a few well-known data representation techniques discussed below

BOW (Bag of Words): The Phenomenon of this technique is to perform a document classification that accepts text input and results in a bag of words. This approach is clear, flexible, simple, and preferred in various ways for feature extraction by discarding grammar and word ordering factors. Entailing these features BOW can be as simple or tough based on the complexity of designing vocabulary of known words and summarizing their presence. This model emphasizes using the frequency of each word as a feature for classifier training. Its application is mainly in areas of Machine Learning[9] where common value depictions are available, like binary features values and frequency values. But a constraint element on ML is that they do not work with raw text but need to be transformed into vectors of numbers.

Natural Language Processing and Biomedical Concept Representation: NLP, a combination of Artificial Intelligence and linguistics which helps in extracting essence rapidly with a huge amount of textual data, is an intelligent analyst of written languages. It excels in areas like information extraction deducing structured data from a given text. This feature of NLP is considered in our work for pulling semantic relationships between data attributes. Our ML approach works with NLP resulting in better representation.[10] The POS tagger used in the present work which executes on the entire dataset to extract vital features like phrases related to nouns, verbs, and preferably biomedical concepts from phrases within datasets.

Medical Concept Representation (UMLS): Unified Medical Language system by virtue works with more general feature representation. This is a knowledge repository from US National Library for medicine. It aims at integrating and distributing medical terminology classifying coding framework and related sources resulting in an effective and efficient Interoperable biomedical system. The three knowledge sources available are Meta thesaurus, semantic network, and lexicon lexical tools. Our work relies on mapping medical features like diseases, treatment, cure, prevent with their respective UMLS terms. Ontology extraction takes its place in Medical Semantic Annotation to extract synonyms, definition, and deduce categories from the sentences in the medical records.

A deep insight into classification and representation also raises a query like how input

needs to be served for these techniques. A basic method is to preprocessing the input and directing it to further classification. Our focus is on PCA transformation which well pairs with Machine learning techniques resulting in an enhanced and enriched Accuracy.

PCA: Always the opted choice in Machine Learning for feature reduction suffers from primary complaint i.e. high dimensionality[11] with results in the model over fitting in ML. This decreases the capacity of the system to generalize examples other than the training set.[12] PCA, an unsupervised statistical technique helps in modeling systems with reduced features boosting learning rates, and reducing[13] computational overhead by eliminating duplicate features. It is a good choice for image compression and shows its real color in an Image processing environment. PCA moves on with few assumptions like a sample size of minimum 150 cases, features sets are co-related and outlier free, Maintaining Linear property, etc. for better operation. During the Workflow of PCA[14], we start with normalizing the data, constructing a covariance matrix that defines all possible relationships among various dimensions for Eigen decomposition, and finally selects the optimal count of Principal components. Many Applications of PCA in neural networks proved to exhibit improved efficiency by converging over fitting problem to its minimum. Despite these nourishing features it also rings its danger bells in few aspects like outliers, reducing Independent feature recognition i.e. Interpretation, lowering performance in cases of lowered feature co-relation (less co-related features) down-trending accuracy of the classifier when handling discriminating class characteristics.

Proposed Model

Our proposed model promises enhanced results as the approach entails data preprocessing in reducing the dimensionality of considered multi variant dataset by projecting it on to a lower dimensional space thereby gearing up the speed of classification algorithm. The transformation process proceeds by dropping irrelevant features resulting in accelerating the analysis process and proving it to be efficient when compared against algorithms considered without such transformations and reductions. The existing approaches considered the datasets as extracted from sources and subjected them to the classification process with results reflecting a far trade off in comparison with proposed. Several profound quality parameters that showed improvement of proposed framework in contrast to SVM, Decision Trees are Accuracy of the classified results, Recall and Precision with so called true positives, false positives and true negatives, and increased time efficiency.

The working of proposed Model is elucidated in 3 phases as follows:

Initially, PCA is used to transform the text which is in the form of sentences in to a huge medical dataset into vector representation for minimizing the number of features. It leads to reduced completely and improved accuracy of the system, which permits ML algorithms to run faster.

The second phase is emphasizing on a predictive model, to find semantic relation like the cure, treatment, diseases, and side effects between disease and treatment by exploring otology, definition, and synonym of texts. It includes removing stop words and using POS tagging with the n-gram technique. The last phase is into encounter classification process from the output obtained from PCA transformation. The CNB classification which is considered optimal classifies is subjected to all the 3 forms of representation i.e. BOW, UML, and NLP. The classification guarantees high probabilities results regardless of the represent assuring that the proposed method excels in accuracy, f-measure, recall, and precision.

This section describes the procedure of the proposed approach that comprises short text related sentence processing, medical concept semantic attribute relation, exploration of semantic relations, and inference data search from medical data sources is described.

Algorithm for Proposed Model

Step by step Algorithm procedure to explore semantic relations present in short text disease treatment sentences on transformed dataset

i/p:text in the form of sentences

o/p: Classification of semantic relations with medication sentences

1 For each word present in text sentence T
2 *α*add each word to text //α :list of words in given sentences;
3 End for loop with last sentences
4 Related semantic Words i.e. {".","?","!","!!","!!!","??","?!","!?"}
5 // Word tokenization starts
6 For each token in α

7 If(token= Related semantic Words)
8 *β.add(sentence)* // β:array holding semantic words;
9 Sentence.empty ()
10 Else
11 Sentence.add(token)
12 End
13 //Remove irrelevant words(i.e.stop words, PoS: Parts of speech tagger)
14 stopWords('without','also','must','might')
15 for each sentence in β
16 if (words present in sentence)
17 remove all the words from sentences
18 end if
19 End for
20 β_s*.add(sentence(with(referral(words))))*
21 End for
22 Start PoS
23 $_{\beta.}$ *add(POSTagger(sentence))*
24 Semantic relation of(POSTagger)
25 SRof (sentence)
26 Classification Starts
27 *Tui = get Semantic Type* (*sentence Concept, df: type Of Semantic,? any*)
28 *if(getSemanticGropWithSimilarRelations(tui,calb c:inGroup,Semant icRelationType)==true*
29 Else
30 *tuiSyn=get Semantic Relation Type (sentence Concept, synonyms,df:type ,?any)*
31 If *get Semantic Relation Type (tui, Anc; calbc; Semantic Type)==true*
32 Semantic Relation=sentenceConcept
33 Else
34 didn't sentence related concept
35 End if
36 End if
37 End for

Performance Analysis

Recall: An important measure for focusing actual positives among scenarios where our evaluation choice is to capture more positives from datasets. Recall is formulated as

$$Recall\ (R) = \frac{Tp}{Tp+Fn}$$

Precision: A choice metric to answer the question The Performance of our Proposed Multi- variant Approach is evaluated against other well-known classification algorithms such as SVM, DT, and CNB with metrics like Accuracy, Recall, Precision, and Time Efficiency. These metrics help in measuring and evaluating the varying sequences of results.

Accuracy: A quintessential metric of Classification well performs for binary and multiclass is the number of true results for total cases being classified. The computational formula defining Accuracy measure is defined below.

$$Accuracy = \frac{(Tn + Tp)}{(Tn + Tp + Fn + Fp)}$$

where *Tn*, *Tp* are true positives, true negatives and *Fn, Fp* are false positives and false negatives respectively. of the correctness of the predictions made during the classification process. It can be computed using the formula defined below.

$$Precision(P) = \frac{Tp}{Tp+Fp}$$

where *Tp* and *Fp* holding true positives and true negatives of classification .

Time Efficiency: This metric measure the time lapsed for attaining efficient results by the classification algorithms when simulated for 30 sec for JSIM.

In our experimental setup, we have taken a Pub med medical dataset containing published records of various diseases. Among them datasets related to Arthritis are considered as Data set1, Cancer related records are considered as Data set 2, Diabetics as Data set 3, Asthma as Data set 4, and Alzheimer's records are taken as Data set 5 in our table of values.

Observations

In Fig.1 an improved accuracy of proposed model when compared with other classification algorithms like CNB, SVM and DT is projected. The parameter

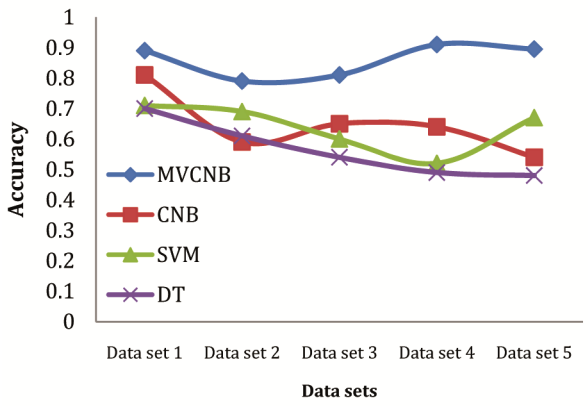

Fig. 1 — Accuracy of MVCNB Vs DT, SVM, & CNB

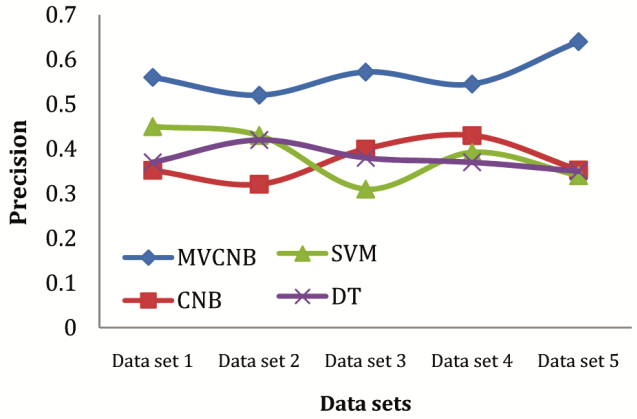

Fig. 2 — Precision of MVCNB Vs DT, SVM, & CNB

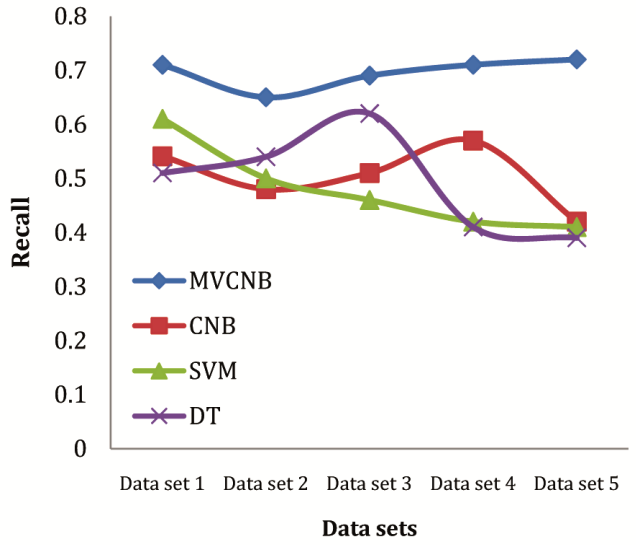

Fig. 3 — Recall of MVCNB Vs DT, SVM, & CNB

depicts a high raise in accuracy with a value of 0.89, with a minimum improvement of 0.07 and more.

Respectively, Fig. 2 shows Precision projecting higher degree of correctness represented with blue spikes by MVCNB with 0.56 in contrast to other considered SVM, DT and CB with high variance of minimum 0.1 that gradually promises to leverage with increased data set size. The visual improvement of Recall value 0.72 for MVCNB in contrast to minimum value for DT when compared with positives of several algorithms and data set sizes are depicted Fig. 3.

Finally, Fig. 4 depicts MVCNB reflecting higher Time efficiency of value 0.71 in maximum cases and few others like SVM with 0.54 doing well at times for some data sets.

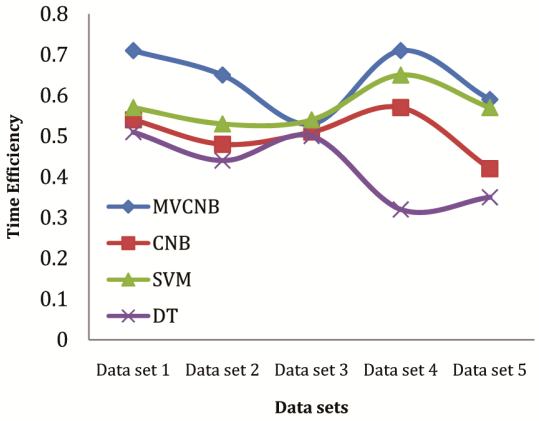

Fig. 4 — Time Efficiency of MVCNB Vs DT, SVM & CNB

Conclusions

Referring to the above results and observations, it is possible to claim few assertions that domain-specific information improves the data mining task consequences. Among the various classification Models, the probabilistic models turn out to be consistent and reliable choice for medical domains holding short text. Our task of semantic relation identification was accomplished with ML-based techniques further subjecting it to PCA transformation on medical datasets. The transformed data succeeds in projecting to a lower dimensional space by eliminating irrelevant features thereby paving path for improved performance. The pre-processed data ensures effective analysis of disease treatment in medical datasets with CNB classification, resulting in improvement of evaluation parameters like accuracy, precision, recall, and improved execution time over other competitive classification algorithms. As a note of further scope, medical data undergoing intense challenges during publishing need to be in fusion with evolving technologies for attaining improvised results. Further the scope can be extended by identifying Drug similarities with predictive models for faster treatment and Optimized approaches for Drug repositioning. Such advancements in clinical domain leverage the recovery rate of patients, benefitting the society and imaging the essence of research.

References

1 AlGhunaim S & Al-Baity H, On the scalability of machine-learning algorithms for breast cancer prediction in big data context, *IEEE Access*, **7** (2019) 91535–91546, doi: 10.1109/ACCESS.2019.2927080.

2 Dehkordi S K & Sajedi H, Prediction of disease based on prescription using data mining methods, *Health Technol*, **9** (2019) 37–44.

3 Frunza O, Inkpen D & Tran T, A machine learning approach for identifying disease-treatment relations in short texts, *IEEE Trans Knowl Data Eng*, **23(6)** (2011) 801–814.

4 Sisodia D & Sisodia D S, Prediction of diabetes using classification algorithms, *Procedia Comput Sci*, **132** (2018) 1578–1585.

5 Xing W & Bei Y, Medical health big data classification based on KNN classification algorithm, *IEEE Access*, **8** (2020) 28808–28819. doi:10.1109/access.2019.2955754.

6 Cao J, Lv G, Chang C & Li H, A feature selection based serial SVM ensemble classifier, *IEEE Access*, **7** (2019) 144516–144523, doi: 10.1109/ACCESS.2019.2917310.

7 Ismail W N, Hassan M M, Alsalamah H A & Fortino G, CNN-based health model for regular health factors analysis in internet-of-medical things environment, *IEEE Access*, **8** (2020) 52541–52549, doi: 10.1109/ACCESS.2020.2980938.

8 Muzaffar A W, Azam F & Qamar U, A relation extraction framework for biomedical text using hybrid feature set, *Comput Math Methods Med (CMM)*, (2015), Article ID 910423, https://doi.org/10.1155/2015/910423.

9 Kohavi R & F Provost, Glossary of terms, machine learning, *Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process*, **30** (1998) 271–274.

10 Ledesma S, Ibarra-Manzano M-A, Cabal-Yepez E, Almanza-Ojeda D-L & Avina-Cervantes J-G, Analysis of data sets with learning conflicts for Machine Learning, *IEEE Access*, **6** (2018) 45062–45070 doi: 10.1109/ACCESS.2018.2865135.

11 Reddy G T, Reddy M P K, Lakshmanna K, Kaluri R, Rajput D S, Srivastava G & Baker T, Analysis of dimensionality reduction techniques on big data, *IEEE Access*, **8** (2020) 54776–54788.

12 Choubey D K, Kumar P, Tripathi S & Kumar S, Performance evaluation of classification methods with PCA and PSO for diabetes, *Netw Model Anal Health Inform Bioinforma* **9**(**5**) (2020). *https://doi.org/10.1007/s13721-019-0210-8*.

13 Hassler A P, Menasalvas E, García-García F J, Rodríguez-Mañas L & Holzinger A, Importance of medical data preprocessing in predictive modeling and risk factor discovery for the frailty syndrome, *BMC Med Inform Decis Mak* **19(33)** (2019). https://doi.org/10.1186/s12911-019-0747-6.

14 H Rupa & T Asha, A linear model based on principal component analysis for disease prediction, *IEEE Access*, **7** (2019) 105314–105318.