International Journal of Engineering & Technology, 7 (2.24) (2018) 11-15

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Disease Risk Prediction using SVM based on Geographical Location

Abarna A. R[1*], A. Umamakeswari [2]

[1]M.Tech., Computer Science Engineering, [2]Associate Dean-CSE
School of Computing, SASTRA Deemed to be University, Thanjavur-613401
**Corresponding Author E-mail: arabarna247@gmail.com*

Abstract

Nowadays, people get more useful information from the internet and other technological platforms like social media. Plenty of health-care related information is available in social media where people spend more time in it. The existing methodology doesn't include location in particular the user similarity based on the attributes. The proposed method specifies the assessment of disease risk by Support Vector Machine (SVM) algorithm to identify the similarity between the users based on the geographical location and then recommends the health expert to the users. This method also identifies the fake users and validates them. The health-care associated with big-data can be performed effectively in the proposed framework. The experimental output shows that the proposed method is more effective when compared with Collaborative Filtering based Disease Risk Assessment.

Keywords: *Support Vector Machine, Geo location, Health care, Big Data, Disease Prediction.*

1. Introduction

The increase in mobile computing devices plays a vital role over the Internet. Huge data volumes have the most challenging task of managing the data from many different sources, for that, they need big data licensed devices and techniques. Big data describes the data with eminent volumes, eminent variety, and high velocity [1]. Characteristics of big data are mentioned in [13]. In the current scenario, rapid improvement of data has been noticed in the healthcare-related domain as in digital commerce [2].

Big data techniques are mostly used in healthcare and biomedical which originates from genomic-driven and payer-provider sources. Genomic-driven includes genotyping, sequencing data for next generation and gene expression data. Payer-provider consists of Electronic Health Record (EHR), patient's feedback, insurance records, and pharmacy prescription [14]. The need in transposing and integrating the computerized medical information was distributed across many areas like providers of health-related insurance, care, medical research centres, and laboratories. For exchanging infrastructure, it requires powerful, robust and outlays productive storage ideas. The proposed structure facilitates the patient to use healthcare related information at no cost.

The existing framework offers two types of services (1) disease risk prediction and (2) health expert recommendation service. CFDRA (Collaborative Filtering-based Disease Risk Assessment) is used to achieve the assessment of the disease risk. This CF method is used to predict the user's disease by comparing the current user with querying users [15]. Age, height, gender, weight, and family history are the attributes which are used to compare between current and querying users. The recommendation of health expert from Twitter module suggests experts to the users. The expert may be a doctor or non-doctor to help the enquiring users in seeking advice from health expert at free cost.

In this proposed method, SVM algorithm is used to identify similar users between the current and existing users based on the geographical location. The SVM model is a supervised machine learning technique which was based on statistical theory. It was first suggested by Cortes and Vapnik in 1995. It can be used for classification as well as regression. SVM works on non-linear and linear feature. The non-linear method is function-independent whereas the linear method is used as a separator for two different data to determine two unique classes in the multidimensional settings [3].

Section 2 explains the related work, proposed methodology is given in Section 3, and Section 4 discusses the results and analysis of proposed method.

2. Related Works

In social media, the healthcare communication shows a dramatic change with incredible speed in the healthcare industry. Healthcare analytics database is growing exponentially, which can cater to self-education at individual-level, and it can be used by patients [15]. This method is most efficient by using Naïve Bayes to predict the patients from their heart disease followed by Decision Tree and Neural Network. The possibilities of a patient getting heart attack can be predicted by some medically related profiles like age, height, sex, weight, pre-diabetes, and blood pressure. Naive Bayes method is better than the decision tree to identify the importance of medical predictors [16].

To seek the advice, users can communicate with the health experts through Twitter. Those services are fully utilized by the users provided by the health communities through online at no cost. Hyperlink-Induced Topic Search (HITS) methodology is used to identify the health expert candidate based on the keywords related to health which is frequently tweeted on the Twitter. HITS approach is also called as hubs and authority model [17][21]. The most

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

reliable sub-set of fraudulent user accounts were identified, and it is used as a valuable tool for own data convention and dissemination. Whether the Twitter account of the user is legitimate or malicious can be identified by two techniques namely, honeypot and machine learning technique. To attract the spammers honeypot is used, so that, their information can be easily retrieved. Then, the spammer's activities have been understood by analyzing machine learning approach [18]. Recommended system is a powerful technology for extracting secondary value. It uses a new algorithm for Collaborative Filtering (CF) based recommendation system. The result produces high-quality recommendation and large datasets [19]. The Marketing Decision Support System (MDSS) method is developed based on the Multilayer Perceptron (MLP) to support analysis of heart disease. MLP includes three different layers namely, an input layer, output layer, and hidden layer. The input layer comprises of 40 types of inputs; it is partitioned into four different groups and then coded by proposed algorithms. With the help of cascade learning method, the number of variables in the hidden layer is obtained. The nodes presented in the output layer are correlated to any one of the heart diseases. The result is of high accuracy [4].

Assad Abbas et al. [5]employed a Collaborative Filtering approach for finding similarity between current and existing users based on the user's profile. This approach provides high accuracy for determining the similarity, and it is handled by Recommender System (RS). However, the drawback of the collaborative filtering method is time-consumption. But the proposed methodology uses Support Vector Machine to assess disease risk based on the geographical location.

Chakraborty et al.[6] used Random Forest (RF) approach for predicting disease with the help of HCUP datasets. It is based on the past medical diagnosis of the individual user. The RF method is associated with sub-sampling. The advantages of this approach are, the HCUP dataset would predict eight types of disease and also achieve high accuracy of prediction. The restriction of the RF method has the problem of overfitting with disordered datasets.

Khatib et al. [20] used fuzzy set logic to identify the disease risk associated with the heart, i.e., coronary thrombosis. This logic provides personalized suggestion to reduce the risk level of disease. The disadvantage of this logic is to restrict manipulating diversification of healthcare related information. The author [16] employed Naïve Bayes method for Alzheimer Disease (AD) to determine risk assessment by genomic-driven bigdata.

3. Proposed Method

The proposed methodology introduces assessment of disease risk and recommendation of health experts to the user on Twitter. The users are validated while enquiring about their disease as fraudulent or valid. This can be identified using an attribute of the users like user's mail id, mobile number, etc. The proposed system uses Support Vector Machine (SVM) algorithm for classification.

The SVM method used collaborative filtering technique for a recommendation which is known as SVMCF4SR[7]. It recommends efficient web services to the users, where it detaches the hyperplane from the most historical data, the customer may not favor these separating services. And also, this SVMCF4SR method is appropriate to calculate the desired degree of the recommended services. The degree of a user can be determined precisely with the scope of the services and hyperplane points. Coronary Artery Disease (CAD) patients have been automatically analyzed by Support Vector Machine (SVM) classifier which is used to segregate the two different classes of data. SVM method increases the accuracy of CAD diagnosis[8].

The classification of SVM model can also be used in medically related datasets [3]. Padmavathi Janardhanan.et al. ensures that the performance of SVM classification model is more efficient than the other techniques namely, Naïve Bayes classification and RBF (Radial Basis Function). They used various types of medically related datasets, compared the result of prediction and recorded it. This dataset includes three kinds of diseases called cancer dataset, heart and diabetes dataset. The performance has been done in WEKA platform, and the determined result says that the SVM model is the most efficient and robust classifier for medically related datasets.

In this paper, the proposed system is implemented by Supervised Learning algorithm or Support Vector Machine classifier. This algorithm is used to determine the similarity between enquiring users and existing users in geographical location. For that, the first need is to collect the disease datasets which consist of symptoms of every disease and also the medicine name for the corresponding disease. The symptoms of flu diseases are mentioned in Fig 1.

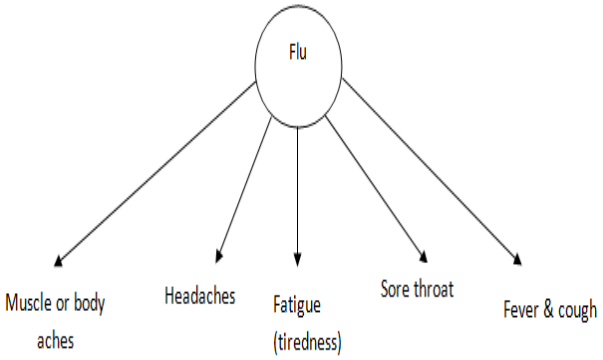

Fig. 1: Symptoms of Flu

The similarity among the users can be determined based on the symptoms which are tweeted by the inquiring user. A user can tweet one or more tweets, for every tweet the supervised learning algorithm identifies the disease. The stop words remove the unwanted words in the tweets. After filtering the unwanted words, it detects the similarity between users by using symptoms. Based on the symptoms, the disease name is displayed along with the drug name for that disease. This algorithm is also used to suggest the specialist according to user's disease. This algorithm also explains how the user can chat with that specialist to know about the risk of their disease as shown in Fig 2.

The main advantage of recommending the specialist to the user is to consult doctor at no cost. However, the supervised learning mechanism is used to decrease the time consumption of finding similar users based on the geolocation. Table 1 explains various symbols used in this paper.

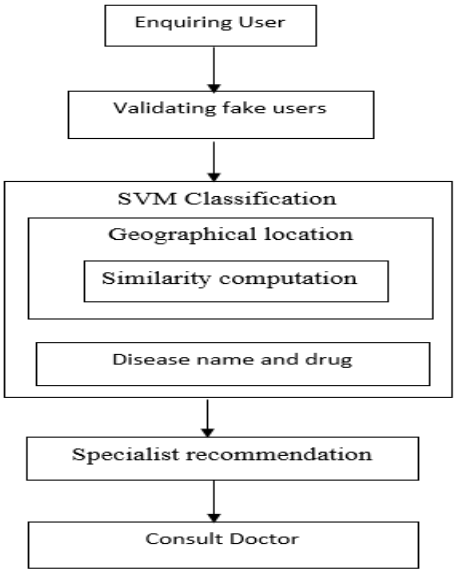

Fig. 2: Proposed System Architecture

Table 1: Symbols and descriptions

S. No	Symbols	Descriptions
1	A	Set of attributes
2	a	Attribute
3	P	Profile of users
4	EP	Profile of existing users
5	e	Existing user
6	S	Set of all symptoms in profile of users
7	s	Symptoms of particular user
8	Q	Set of enquiring users
9	qu	Enquiring user
10	G	Set of geographical location of users
11	g	Geographical location of particular users
12	Sp	Shortlisted profile of users

Algorithm:

Input: Enquiring user q for disease d
Output: Name of the disease along with the experts.

1: for attribute a ∈ A do
2: A← getAttribute (u)
3: P ← retrieveProfiles (A)
4: end for
5: for s ∈S do
6: s ← getSymptoms (P)
7: end for
8: for Enquiring user qu ∈ Q do
9: if (qu == true) then
10: Sp ← {s ∈ P}
11: else
12: Sp ← {s ∉ P}
13: end for
14: for geoLocation g ∈ G do
15: if (g==true && q== true)
16: Sp ← {s ∈ P}
17: res ← retriveDisease (P)
18: else
19: Sp ← {s ∉ P}
20: end for
21: recommend expert(res)

In line 1-4, the profile of an inquiring user can be retrieved along with an attribute of the users. In line 5-7, symptoms of the users are recovered. Line 8-Line 13 shortlisted profile of a user is fetched by comparing symptoms of the user profile with the enquiring user symptoms. Line 14-Line 20 performs symptom comparison based on the geographical location. When the disease is retrieved, then the symptoms of the user within the geolocation might be identified. Line 21 is used to recommend the expert based on the disease.

4. Experimental Results

The supervised learning methodology is used to determine the effectiveness of the proposed framework by comparing different types of popular approaches such as collaborative filtering, CART, Naïve Bayes, Logistic Regression and MLP[9]. The brief explanation of the above methods is as follows:

4.1 Collaborative Filtering Approach:

RS (Recommender System) method uses Collaborative Filtering Algorithm. It has two types of functions such as narrow function and most general function. Automatic prediction of the interest of the user is made by the narrower function which collects the data from several users [5].

4.2 CART:

Classification and Regression Tree is a classification model based tree. It uses cross-validation to prefer convenient tree. The classification tree is defined as target values which are taken from any finite set of variables. In regression tree, the targeted values are taken from continuous variables. For the classification and regression purpose, this model is used as prediction techniques in healthcare.

4.3 Naive Bayes classifier:

Naive Bayes classifier uses the independent attribute [16]. This classifier is used to establish the high capability of prediction. That is the presence of any other feature is unrelated to the presence of a particular feature in a class. Conditional probability using Naive Bayes theorem can be written as given in equation (1).

$$P(C \mid X) = \frac{P(X|C)P(C)}{P(X)}$$

Where P(C|X) - Posterior Probability
P(X|C) - Likelihood
P(C) - Class Prior Probability
P(X) - Predictor Prior Probability

4.4 Logistic Regression:

Logistic regression is a standard regression and classification method for predicting risk [10]. The issue of the logical regression depends on different projections or features. DV (Dependent variable) is categorical in logistic regression. It can take a binary value like '0' and '1', which produce an outcome such as, true/false or pass/fail or win/loss.

4.5 MLP:

MLP-Multilayer Perceptron is a class of neural network. It is widely used in decision support for medically related searches. An analysis of disease based on the heart can be supported by developing a decision support method based on MLP. MLP incorporate with three layers such as a hidden layer, input, and output layers. Inputs and outputs are received at input and output layer respectively, and the process is performed at the hidden layer. MLP also used in CMS(Cellular Manufacturing System) [11].
Some of the standard evaluation metrics were used to estimate the performance of SVM method with other techniques such as precision, F-measure, and recall [12].
The precision is also known as positive predictive value. It is defined as the ratio between TP (True Positive) values and the sum of the instance of defects. The precision is given in equation (2).

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

Where TP is a probability of hits,
FP is a probability of fail and hit.

Recall is the ratio between the precisely identified defects and the sum of the instance of defects. It is also referred as sensitivity and it is expressed as given in equation (3).

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

Where FN is a probability of rejections.

The F-measure harmonic mean is the combination of precision and recall, and the equation can be written in the equation (4).

$$\text{F-measure} = \frac{2TP}{2TP+FP+FN} \quad (4)$$

Among these three methods, F-measures classifier is the most effective model because it combines other two methods.

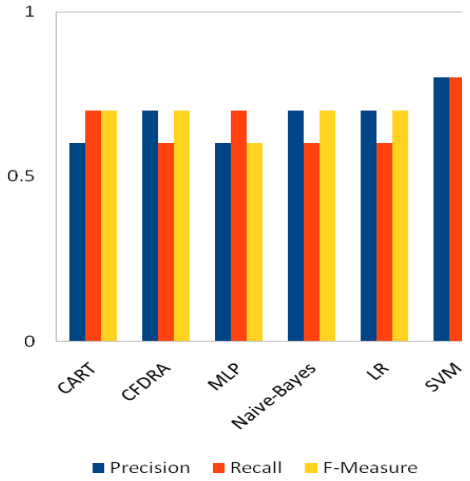

Fig. 3: Comparison of proposed method SVM with other related techniques for case (YES)

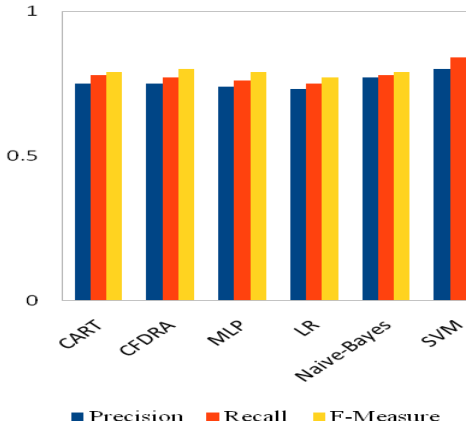

Fig. 4: Comparison of proposed method SVM classifier with other related techniques for case (NO)

These approaches were estimated by verifying the accuracy of the symptoms name which is specified by the user. For example, 'ever diagnosed flu' is the disease in the database it might be "YES" or "NO". The "YES" or "NO" represents whether the patient is affected by flu or not.

Fig 3 represents the comparison between the outcomes, for the case "YES". The case YES denotes, the patients who had flu disease. In Fig 4, the comparison of outcome for the case "NO" represents the testing patient should not contain flu disease.

The comparison results indicate that the proposed method is the best technique to identify the disease using symptoms on geographical location. The outcome reduces the time consumption of finding similarity, and it is highly scalable method when compared to all the techniques mentioned above.

5. Conclusion

The existing methodology failed to validate fake users and also to find the similarity in user symptoms based on the particular location. The CFDRA method consumes more time to identify the similarity of users. In the proposed method, fraudulent users can be identified by their attributes. An algorithm called Supervised Learning has been used to identify the similarities between two or more users based on the geographical location, which then recommends the experts of that disease to the querying users at free of cost. The disease risk prediction results are compared with various regressions and classifiers such as collaborative Filtering, MLP (Multi-Layer Perceptron), CART (Classification and Regression Tree), Naive Bayes, and logistic regression. The result obtained is highly scalable, less time consuming and produces a more efficient model for recommending the health expert to the users.

References

[1] Ishwarappa and J. Anuradha, "A brief introduction on big data 5Vs characteristics and hadoop technology", Procedia Comput. Sci., vol. 48, C, pp. 319–324, 2015.

[2] A. Abbas, K. Bilal, L. Zhang, and S. U. Khan, "A cloud based health insurance plan recommendation system: A user centered approach", Futur. Gener. Comput. Syst., vol. 43–44, pp. 99–109, 2015.

[3] P. Janardhanan, L. Heena, and F. Sabika, "Effectiveness of support vector machines in medical data mining", J. Commun. Softw. Syst., vol. 11, no. 1, pp. 25–30, 2015.

[4] K. Zhao, J. Yen, G. Greer, B. Qiu, P. Mitra, and K. Portier, "Finding influential users of online health communities: a new metric based on sentiment influence", J. Am. Med. Inform. Assoc., vol. 21, no.2, 2014.

[5] A. Abbas, M. Ali, M. U. Shahid Khan, and S. U. Khan, "Personalized healthcare cloud services for disease risk assessment and wellness management using social media", Pervasive Mob. Comput., vol. 28, pp. 81–99, 2016.

[6] M. Khalilia, S. Chakraborty, and and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest", BMC Med. Inform. Decis. Mak., vol. 11, no. 1, 2011.

[7] L. Ren and W. Wang, "An SVM-based collaborative filtering approach for Top-N web services recommendation", Futur. Gener. Comput. Syst., vol. 78, pp. 531–543, 2018.

[8] ated diagnosis of coronary artery disease (CAD) patients using optimized SVM", Comput. Methods Programs Biomed., vol. 138, pp. 117–126, 2017.

[9] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury, "Application of support vector machine modeling for prediction of common diseases: The case of diabetes and pre-diabetes", BMC Med. Inform. Decis. Mak., vol. 10, no. 1, 2010.

[10] J. Shang, M. Chen, H. Ji, D. Zhou, H. Zhang, and M. Li, "Dominant trend based logistic regression for fault diagnosis in nonstationary processes", Control Eng. Pract., vol. 66, January, pp. 156–168, 2017.

[11] A. Delgoshaei and C. Gomes, "A multi-layer perceptron for scheduling cellular manufacturing systems in the presence of unreliable machines and uncertain cost", Appl. Soft Comput. J., vol. 49, pp. 27–55, 2016.

[12] A. Abbas, M. U. S. Khan, M. Ali, S. U. Khan, and L. T. Yang, "A cloud based framework for identification of influential health experts from twitter", IEEE 15th Int. Conf. Scalable Comput. Commun. 20, pp. 831–838, 2016.

[13] M. Hilbert, "Big Data for Development", Dev. Policy Rev. Overseas Dev. Inst., vol. 1, no. 1, pp. 2012–2013, 2013.

[14] S. Chung Chow, "On Big-Data Analytics in Biomedical Research", J. Biom. Biostat., vol. 6, no. 3, 2015.

[15] A. Kotov and A. Kotov, "Social Media Analytics for Healthcare Chapter 1 Social Media Analytics for Healthcare", 2015.

[16] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques", 2008 IEEE/ACS Int. Conf. Comput. Syst. Appl., pp. 108–115, 2008.

[17] Dhivyalakshmi, S. Umamakeswari, A. "The role of big data analytics in hospital management system", International Journal of Pure and Applied Mathematics, vol.115, no. 7 Special Issue, 2017, Pages 31-35.

[18] S. Gurajala, J. S. White, B. Hudson, and J. N. Matthews, "Fake Twitter Accounts: Profile Characteristics Obtained Using an Activity-based Pattern Detection Approach", Proc. 2015 Int. Conf. Soc. Media Soc., p. 9:1--9:7, 2015.

[19] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl, "Item-based collaborative filtering recommendation algorithms", Proc. tenth Int. Conf. World Wide Web - WWW '01, pp. 285–295, 2001.

[20] E. F. Erkan, Ç. Teke, and D. Güleryüz, "A Fuzzy Expert System for Risk Self-Assessment of Chronic Diseases", vol. 18, no. 6, pp. 29–33, 2016.

[21] Dhivyalakshmi, S. Umamakeswari, A. Aishwarya, Subramanian, E.R.S. Gurubaran, B. Shailesh Dheep, G., "Data analytics for smart HMIS", International Journal of Pure and Applied Mathematics, vol.115, no. 7 Special Issue, 2017, Pages 167-173.

[22] T. Padmapriya and V. Saminadan, "Inter-cell Load Balancing technique for multi-class traffic in MIMO-LTE-A Networks",

International Journal of Electrical, Electronics and Data Communication (IJEEDC), ISSN: 2320- 2084, vol.3, no.8, pp. 22-26, Aug 2015.

[23] S.V.Manikanthan and D.Sugandhi "Interference Alignment Techniques For Mimo Multicell Based On Relay Interference Broadcast Channel, "International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE) ISSN: 0976-1353 Volume- 7, Issue 1 –MARCH 2014.