International Journal of Recent Engineering Research and Development (IJRERD)
ISSN: 2455-8761
www.ijrerd.com || Volume 03 – Issue 02 || February 2018 || PP. 34-38

Heart Attack Detection Using Bag of Words with Machine Learning: A Review

[1]Surbhi.Kadu, [2]Dr.G.R.Bamnote
ME student Dept. of CSE,
PRMIT&R Badnera, Amravati, India,

Abstract: According to recent survey by WHO organization 17.5 million people dead each year. It will increase to 75 million in the year 2030. Medical professionals working in the field of heart disease have their own limitation, they can predict chance of heart attack up to 67% accuracy, with the current epidemic scenario doctors need a support system for more accurate prediction of heart disease. Machine learning algorithm opens new door opportunities for precise predication of heart attack. The Bag of Words model (BOW) originated in natural language processing. It makes the simplifying assumption that the order of the words in a sentence or text document is of negligible importance for classifying it. This paper highlights the important role played by the machine learning algorithm in analyzing huge volumes of healthcare related data in prediction and diagnosis of disease.

Keywords: Heart Diseases, Data mining, k-mean clustering, bag of words

1. Introduction

Healthcare organizations are facing with challenges to give cost-effective and high quality patient care. Both administrators and clinicians need to analyze a wealth of data available in the databases of healthcare information systems in order to discover knowledge and to make informed decisions. This is basic specifically to improve the viability of sickness treatment and preventions. It becomes of more important in case of heart disease (HD) that is regarded as the primary reason behind death in adults. Data mining serves as an analysis tool to discover unapparent or hidden relationships and patterns in HD medical.

There are five models constructed of single and combined data mining techniques to support clinical decisions in (HD) diagnosis and prediction. The five systems give automatic pattern recognition and attempt to reveal relationships among different parameters and symptoms of HD. Each system exhibits set of strengths and limitations in terms of the type of data it handles, accuracy, ease of interpretation, reliability and generalization ability. Weak generalization ability is still a big open issue for data mining in healthcare mainly because of the lack of input data and cost of re-processing. Data can be a great asset to healthcare organizations, but they have to be first transformed into information". More demands are placed on using this information to build knowledge that enables the strategy of healthcare organizations: maximize patient care and minimize cost. The healthcare environment is perceived as being "information rich" but "knowledge poor"

There is a wealth of clinical and administrative data available within healthcare systems, however, there is a lack of effectual analysis tools to discover knowledge contained in the databases of these systems . Knowledge Discovery in database (KDD) refers to the "non trivial extraction of implicit previously unknown and potentially useful information about data". Data mining is the core of KDD and which is defined as "a process of selection, exploration and modeling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results for the owner of database". The discovered knowledge in healthcare databases can be used by healthcare administrators to improve operations and quality of service. It can be also used by healthcare professionals to improve their medical practice and patient care. The objective is to provide an appraisal of current state-of-the art applications of knowledge discovery in medical databases using data mining techniques to predict heart conditions. There are several of applications that use a single or combination of predictive data mining techniques to improve prediction accuracy. The bag of words technique using k-mean clustering and machine learning to make predictions over available medical data.

A bag-of-words model, or BOW for short is a way of extracting features from text for use in modeling, such as with machine learning algorithm. The approach is very simple and flexible, and can be used in a myriad of ways for extracting features from documents. A bag-of-word is a representation of text that describes the occurrence of words within a document. It is called bag-of-words because any information about the order or structure of words in the document is discarded. The model is only concerned with whether known words occur in the document, not where in the document. The bag-of-word model is a way of representing text data when modeling text with machine learning algorithms. The bag-of-model is simple to understand and implement and has seen great success in problem such as language modeling and document classification.

2. Related Work

Asha Rajkumar [11] suggested a prototype of IHDPS (Intelligent Heart Disease Prediction System) implementing data mining algorithms, like Naive-Bayes, Decision Trees and Neural Network. The final output of these algorithm describe that each method has its different capabilities in the purpose of the described mining goals. Intelligent Heart Disease Prediction System is simple, easy to use, scalable, expandable and reliable web based predication system that can give output of difficult queries and the traditional decision support system fail to do. In IHDPS author's use the medical profile attributes like age range, gender, high blood pressure and high blood-sugar to discover the symptoms of patients. It is implemented on the .NET platform

Das and Turkoglu [12] introduce a technique that uses Statistical Analysis System (SAS) software version 9.1.3 for investigation of the heart base disease. A neural networks ensemble based techniques generate fresh models by linking the posterior possibilities and discovered values against multiple predecessor models for creating high accurate models. The research exercise were done on the heart diseases dataset to predict heart disease in an entirely automatic way using 3 independent neural networks models to develop the ensemble models. The authors obtained 89.01% classification accuracy, 95.91% specificity and 80.95% sensitivity values on the data drawn from Cleveland heart diseases database.

Srinivas [13] intended application of data mining techniques in discovery of heart diseases. The author used the tanagra tool for implementation of data mining, statistical and machine learning algorithms. Author use the training data set with 14 different attributes and 3000 instances. The experiments were performed on the training data set to measure algorithm performance, in terms of time taken and accuracy. The instances and attributes in the data set were describing the outcomes of various kind of experiments to calculate the efficiency of heart disease. Author divide the data set into two different parts, 30% of data was used as testing and 70% was used as training. The comparison was done on the bases of 10 fold cross validations. The proposed work best performance algorithm results were 52.33% accuracy using Naive Bayes among of these classification algorithms on heart diseases data set.

Shouman[14] presented K-means clustering using the decision tree technique to measure accuracy the heart disease data set. To boost the efficiency of K-means clustering they proposed various kind of centroid selection technique.

Cleveland Clinic Foundation Heart diseases. An accuracy and sensitivity were measured with several centroids selection technique and several bunch of clusters. The ten different runs were performed for the random attribute and random row techniques and the average and best for each technique were measured. Finally author compare the performance of previously used decision tree implemented formerly on the same data with the combine implementation K-mean clustering and decision tree approach. The integrating k-means clustering with decision tree improve resultant efficiency of decision tree to predict heart diseases of patients. The accuracy improved by the enabler technique with 2 clusters was 83.9%. It has been identified from Literature survey, that there are certain issues which are still needs improvement. Classification and association suffers from inefficiency, due to the evidence that it usually produce huge number of rules in associative rule mining. So it very difficult to select best suitable and effective rules from among them. Most of the associative classifiers generate rules in a level wise manner with minimum support pruning. Often that ahead to generation of a large amount of insignificant rules and at the same time good rules with relatively low support are not produced. In most of the case associative classification algorithms support the exhaustive search technique used in the foremost a prior method to finding the rules and need many number of passes over the large data sets. Moreover, they discover frequent items in single phase and create the different rules in different phase exhausting more efforts, storage and processing time.

In case of heart diseases diagnosis the accuracy of heart data set is calculated on the basis of classification methods like Naive Bayes, IBk, Neural Network, Decision tree etc.

Accuracy of correctly classified instances is not sufficient to predict heart diseases on the basis of training data set and need implementation of association classification approach for better accuracy.

3. Proposed Work

A methodology to detect a heart attack on the basis of available patient dataset which contains various symptoms and other health related parameters like B.P, H.R, etc. Various techniques for pre-processing of collected data are applied such as applying stop words removing, stemming, feature reduction and feature selection techniques to fetch the keywords from all the attributes and finally using different classifiers to decide whether their is heart condition or not. Numeric feature representation technique for feature extraction and K-mean clustering algorithm for clustering feature vectors, in various clusters and machine learning technique to make predictive analysis of heart condition.

3.1 Pre-processing:

A single record containing the gathered works of a creator in spite of the fact that are just keen on a solitary work. Or, on the other hand there might be given a vast work separated into volumes (this is the situation for Les Misèrables, as we will see later) where the division into volumes is not imperative to us.

If a long text is break up into (such as a book-length work) smaller chunks so we can get a sense of the variability in an author's writing. If comparing one group of writers to a second group, to sum a particular information about writers belonging to the same group. This will require merging documents or other information that were initially separate. The section illustrates these two common pre-processing step: splitting large texts into smaller "chunks" and aggregating texts together.

Another important pre processing step is tokenization. The process of splitting a text into individual words or sequences of words (*n-grams*). Decisions regarding tokenization will depend on the language(s) being studied and the research question. For example, should the phrase "her father's arm-chair" be tokenized as as ["her", "father", "s", "arm", "chair"] or ["her", "father's", "arm-chair"]. Tokenization patterns that work for one language may not be appropriate for another (What is the appropriate tokenization of "Qu'est-ce que c'est?"?). The segment begins with a substantial discourse of tokenization before covering splitting and merging text.

Tokenizing:

Tokenization when applied to data security , is the process of substituting a sensitive data element with a non-sensitive equivalent, referred to as a token , that has no extrinsic or exploitable meaning or value.

Stemming:

To count inflected forms of a word together. This procedure is referred to as *stemming*. Stemming a German text treat the given words as instances of the word "Wald": "Wald", "Walde", "Wälder", "Wäldern", "Waldes", and "Walds". Analogously, in English the following words would be counted as "forest": "forest", "forests", "forested", "forest's", "forests'". As stemming lowers the number of unique vocabulary items that are needed to be tracked, it speeds up a variety of computational operations. For some kinds of analyses, such as authorship attribution or fine-grained stylistic analyses, stemming may obscure differences among writers. For example, an author might be distinguished by the use of a plural form of a word.

Chunking:

Splitting a long text into small samples is a very common task in text analysis. Many kinds of quantitative text analysis are take as inputs an unordered list of words, breaking a text into small chunks allows one to preserve context that would otherwise be removed; observing two words together in a paragraph-sized chunk of text tells us more about the relationship between those two words than observing two words occurring together in an 100,000 word book. Or, as we will be using a selection of tragedies as our examples, we may consider the difference between knowing that two character names occur in the similar scene versus knowing that those two names occur in the ; same play.

Stopping:

Removal of stop words – Stop words like "and", "the", "of", etc are very common in all English sentences and are not very meaningful in deciding spam or legitimate status, so these words have been taken out from the emails.

Feature extraction process:

Once the dictionary is ready, we can remove word count vector (our feature here) of 3000 dimensions for every email of training set. Each **word count vector** has the frequency of 3000 words in the training file. Of course you might have guessed by now that most of them will be zero. Let us take an example. Suppose we have 500 words in our dictionary. Each word count vector contains the frequency of 500 dictionary words in the training file. Suppose text in training file was "Get the work done, work done" then it will be [0,0,0,0,0,…….0,0,2,0,0,0,……,0,0,1,0,0,…0,0,1,0,0,……2,0,0,0,0,0]. Here, all the word counts are placed at 296th, 359th, 415th, 495th index of 500 length word count vector and the rest are zero.

The python code will create a component vector grid whose lines signify 700 documents of preparing set and sections indicate 3000 expressions of word reference. The value at index '*ij'* will be the number of occurrences of j^{th} word of dictionary in i^{th} file.

K-means clustering :

K-means clustering is a method of vector quantization, originally from signal processing. K-means clustering aims to partition observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a separate of the data space into Voronoi cells. The problem is computationally difficult however, there are efficient heuristic algorithms that are commonly employed and converge quickly to a local optimum .These are usually equal to the expectation-maximization for mixtures of Gaussian distribution via an iterative refinement approach employed by the both algorithms. Additionally, they both use cluster centers to model the data; however, k-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes.

The algorithm has a free relationship to the k-nearest neighbor classifier, a popular machine learning technique for classification that is often confused with *k*-means because of the *k* in the name. One can apply the 1-nearest neighbor classifier on the cluster centres obtained by *k*-means to classify new data into the existing clusters. This is known as nearest centered classified or Rocchio algorithm.

Training the classifiers:

The scikit-learn ML library for training classifiers. It is an open source python ML library which comes bundled in 3rd party distribution anaconda or can be used by separate installation following . Once it is installed, then it is imported in our program.

There are Naive Bayes classifier and Support Vector Machines (SVM). Naive Bayes classifier is a conventional and very popular method for document classification problem. It is a supervised probabilistic classifier based on Bayes theorem assuming independence between each pair of features. SVMs are supervised binary classifiers which are very effective when you have higher number of features. The main goal of SVM is to separate some subset of training data from rest called the support vectors (boundary of separating hyper-plane). The decision function of SVM model that predicts the class of the test data is based on support vectors and makes use of a kernel trick. Once the classifiers are trained, check the performance of the models on test-set.

Conclusion

Heart attack is crucial health problem in human society, so this paper has summarized the pre processing steps and clustering available methods for predication of this disease.

References

[1]. U. Fayyad, G. Piatetsky-Shapiro and P. Smyt, "Data mining to knowledge discovery in databases." American Association for Artificial Intelligence, pp. 37-54, 1996.
[2]. M. Rahama, Data Mining - A search for knowledge, 1st ed., GRIN Verlag, 2012.
[3]. S.H. Liao, P.H. Chu and P.Y. Hsiao. "Data mining techniques and applications - A decade review from 2000 to 2011." Expert Systems with Applications, vol. 39, no.12, pp. 11303-11311, 2012.
[4]. P. Giudici, Applied Data Mining: Statistical Methods for Business and Industry. New York: John Wiley, 2003.
[5]. R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in Proc. of the 20th Int. Conf. on Very Large Data Bases, Santiago, Chile, 1994, pp. 487-499.
[6]. S.R. Ahmed, "Applications of data mining in retail business", IEEE Int. Conf. on Information Technology: Coding and Computing, 2004, vol 2, pp. 455–459.
[7]. O. Kwon and J.M. Sim, "Effects of data set features on the performances of classification algorithms." Expert Systems with Applications, vol. 40, no.4, pp. 1847-1857, 2013.
[8]. J. Singh, H. Singh, A. Kamra, "Recent trends in data mining: A review," in Proc. of 3rd Int. Conf. on Biomedical Engineering and Assistive Technologies, Chandigarh, India, 2014, pp. 138- 144.
[9]. B. Liu, W. Hsu and Y. Ma, "Integrating classification and association rule mining," in Proc. of Knowledge Discovery and Data Mining, National University of Singapore, 1998,
[10]. N. Deepika and K.C. Shekar, "Association rule for classification of heart attack patients." International Journal of Advanced Engineering Science and Technologies, vol. 11, no.2, pp. 253-257, 2011
[11]. Asha Rajkumar, G .Sophia Reena , Diagnosis Of HeartDisease Using Datamining Algorithm, Global Journal of Computer Science and Technology 38 Vol. 10 Issue 10 Ver.1.0 Septembe2010.
[12]. R. Das, I. Turkoglu, and A. Sengur, "Effective diagnosis of 7675-7680, 2009
[13]. K. Srinivas, B.K. Rani and D.A. Govrdhan, "Application of data mining techniques in healthcare and prediction of heart attacks." International Journal on Computer Science and Engineering, vol. 2, no. 2, pp. 250-255, 2011.

[14]. M. Shouman, T. Turner and R. Stocker, "Integrating decision tree and k-means clustering with different initial centroid selection methods in the diagnosis of heart disease patients," in Proc. of Int. Conf. on Data Mining, Australian Defence Force Academy Northcott Drive, Canberra, 2012, pp. 1-7.

[15]. Thabtah and F. Abdeljaber, "A review of associative classification mining." Knowledge Engineering Review, vol. 22, no.1, pp. 37-65, 2007.