Research article Available online www.ijsrr.org ISSN: 2279–0543

International Journal of Scientific Research and Reviews

An Analysis and Implication of Data Mining Techniques using Semi structured Data towards clinical data sets

Suchitra B[1*] and Duraisamy S[2]

[1]Department of Information Technology, Sri Krishna Arts and Science College,CBE-8,India.
[2]Department of Computer Science, Chikkana Government Arts College, Tiruppur, India.
E-mail: suchipradeep@gmail.com, sdsamy.s@gmail.com

ABSTRACT

Data mining plays a dynamic role in today's world. In this era, technological inventions, innovations and also development of algorithms has increasing day by day. One of the prevalent applications of data mining is web mining. Web Mining is an active area. Web structure mining, web content mining and web usage mining are the classifications of Web Mining. This paper stipulates the semi structured data and how data mining techniques are used in the semi-structured data and comparison learning has done to showcase the better technique for supporting semi structured data.

KEYWORDS: Web Mining, Data mining, Support Vector Machine, Polynomial ker

***Corresponding author**

Suchitra B

Department of Information Technology,

Sri Krishna Arts and Science College,CBE-8,India.

E-mail: suchipradeep@gmail.com

INTRODUCTION

Data mining is useful to extract knowledge from databases. Data mining supports different methods and techniques. Data mining area are generally uses classification, Neural Networks, Clustering and Association Rule techniques.

1.1Classification

Classification is an imperative technique of Data mining, here the classification focus on supervised and unsupervised learning. Supervised learning is carried out if samples of data are already having. Unsupervised learning cannot be done like that

1.2 Clustering

Clustering is a process of grouping similar data items or similar objects. Clustering algorithm is classified as Flat algorithms and Hierarchical algorithms. The flat algorithm always starts with random partitioning and refine it iteratively. K –means clustering and model based clustering are the examples of flat algorithms. Bottom up and agglomerative are the hierarchical algorithms. The documents can be used in clustering and they can be categorize as hard and soft clustering. Each document belongs exactly to one cluster is called hard clustering and a document belongs to more than one cluster is called soft clustering

1.3Association Rule

Association rule is based on if/then rules. In association rule, it assumes all data are categorical. Initially association rules are used for market-basket analysis. They use different strategies and data structures the mining exploits sparseness of data, and high minimum support and high minimum confidence values.

REVIEW OF LITERATURE

KNN technique is useful to mine knowledgeable information from medical databases. KNN technique along with genetic algorithm , they propose a new algorithm for effective classification(M.Akhiljabbar*et al*.,2013)[1]

A query enrichment technique is developed to write short query and it maps to intermediate objects (Dou shen et al., 2005)[2]

An automatic classification of new texts that can be improved by a prefiltering the vocabulary to reduce the featured set. They compared artificial neural networks with SVM and they identify the featured set(A.Basu et al, 2002)[3]

A suggestions carried out for ensemble learning and proved it is strong when compared to traditional KNN classifier that uses a different number of neighbors(Ahmad Basheer Hassant et al.,2014)[4]

In order to access the relevant documents in the web, a query based K-Nearest neighbor method is used and it ranks the document on the basis of query rather than customary context based classification (SuneethaManneet al.,2011)[5]

The structured based KNN techniques can be applied to small volume of data and non-structured based KNN techniques can be applied to large volume of data(S.Dhanabalet al.,2011)[6]

A Examination is done on popular semi structured documents mining approach which includes structure and content. In this paper, they compare the documents which include XML and HTML and also compared the semi structured documents content mining approaches include OWL ,XML ,HTML an rdf (Aminamadani et al.,2014)[7]

A possibility of using KNN algorithm with TF-IDF method and framework for text classification is analysed. The framework enables classification according to various parameters, measurements and analysis of result(Bruno Trstenjaket al.,2015)[8]

An approach is suggested to identify web page templates and its tag structure. It focus on sorting semi structured data from web pages and documents. It explores the fetching data as per the requirements using various SQL Queries. (Prashant M. Ahire et al.,2015)[9]

DATA SETS

The clinical Data sets are used in order to find out which algorithm supports semi structured data. The accuracy and time period are given, so that a comparison is made between for the following alogorithm.The datasets used are.

Here we have some attributed Maximum Age, Minimum Age, File Name, Phase, Disease type, gender, and age.

EXPERIMENTS

Data Mining Techniques support Semi structured Data

4.1 K Nearest Neighbors

KNN is an algorithm for stores all the available structured data and classify the new structured data based on measures as distance function. KNN also used in statistical estimation and pattern recognition in non-parametric techniques.

By using a distance function , K nearest neighbor uses a classes that can be classified by a group of neighbors.In this, a instance can be assigned to the class which is most common among KNN. If K = 1, then the instance is assigned to the class of its nearest neighbor.

All three distance measures are only valid for constant variables. The Hamming distance is used for the instance of categorical variables. It focus on theproblem of regularization of the numerical variables between 0 and 1,which contains a mixture of numerical and categorical variables in the dataset.

Selecting the optimum value for K is best done by first reviewing the data. In general, a large K value is more accurate as it decreases the whole noise but there is no guarantee. Cross-validation can also be used to determine a good K value by using an independent dataset to validate the K value. Historically, the optimal K for most datasets has been between 3-10. That produces much better results than 1NN.

4.2 SVM

A Support Vector Machine classifier is a discriminative classifier formally defined by a separating hyper plane. The svm given labelled training data as *supervised learning*, the algorithm outputs an optimal hyper plane which categorizes the structured data in the web. A hyperplane is called as a line in two dimensional spacewhich divides a plane in two parts where in each class lay in either side.

In linear kernel, the equation for prediction for a new input using the dot product between the input (x) and each support vector (xi) is calculated as follows:

$f(x) = B(0) + sum(ai * (x,xi))$

The above equation is used to calculate the inner products of a new input vector (x) with all support vectors in training data. The coefficients such as B0 and ai (for each input) must be estimated from the teaching data by the learning algorithm.

The polynomial kernel can be written as $K(x,xi) = 1 + sum(x * xi)^d$ and exponential as $K(x,xi) = exp(-gamma * sum((x—xi^2))$

RESULTS AND COMPARISON

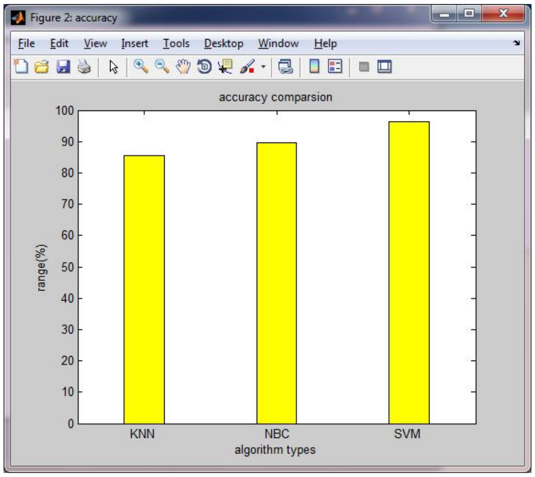

Figure 1: Based on Accuracy comparison for existing (KNN, NBC) proposed (SVM) algorithm

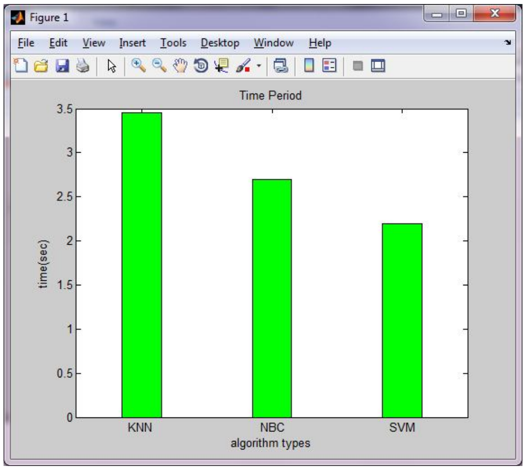

Figure 2: Time period comparison for existing (KNN, NBC) proposed (SVM) algorithm

CONCLUSION AND FUTURE WORK

The Kernel trick is real strength of SVM. It scales relatively well to high dimensional data. The risk of over fitting is less in SVM. The accuracy and time period is less when compared to KNN. The future enhancement focus on developing a tool for the conversion of semistructeddata into structured data, and the accuracy is mainly focused

REFERENCES

1. M. Akhil Jabbar et al. Classification of Heart Disease using K-Nearest Neighbor and Genetic Algorithm. International Conference on Computational Intelligence: Modeling Techniques and Applications.2013; 10: 85-94.
2. Dou Shen, Rong Pan, Jian- Tao Sun. Query-Enrichment for Web-Query Classification.ACM KDDCUP, 2005; 24(3): 320-352
3. A. Basu, C. Watters, M. Shepherd. Support Vector Machines for Text Categorization. 36th International Conference on System Sciences.2002; -5/03;0-7695-1874
4. Ahmad Basheer Hassanat , Mohammed Ali Abbadi and Ghada Awad Altarawneh .Solving the problem of the K Parameter in the KNN Classifier using an Ensemble Learning Approach. International Journal of Computer Science and Information Security. 2014; 12(8).
5. Suneetha Manne, Sita Kumari Kotha and Dr. S. Sameen Fatima. A query based text categorization using K-Nearest Neighbor approach, International Journal of Computer Applications.2011; 32(7).
6. [6].S.Dhanabal and Dr. S. Chandramathi. A review of various K-Nearest Neighbor Query Processing Techniques. International Journal of Computer Applications.2011; 31(7):14-22,
7. Amina Madani et al. Semi structured Documents Mining: A review and Comparison, 17th International Conference in Knowledge based and intelligent information and engineering systems. procedia computer science.2013; 22:330-339.
8. Bruno Trstenjak. KNN with TF-IDF Based Framework for Text Categorization, 24thDaaam International Symposium on Intelligent Manufacturing and Automation. Procediaengineering. 2014; 1356-1364.
9. Prashant M. Atireetal. Extract and Analysis of Semistructured data from websites and documents. International Journal of Computer Science and Mobile Computing. 2015; 2(2) 314-318.