MACHINE LEARNING APPROACH FOR MEDICAL DIAGNOSIS

Manickapriya S. and Nimala K.
Department of Information Technology, SRM University, Chennai, India
E-Mail: priyasara26@gmail.com

ABSTRACT

To overcome clustering problem we use affinity propagation (AP) clustering to handle dynamic data. To handle this, it is important to find the difficulty of incremental affinity propagation (AP) clustering. In AP clustering the newly arrived objects are clustered by adjusting the current data. The message passing concept (MPC) are used for the data communication with each other to produce cluster in parallel, it is used for effective error correction.

Keywords: medical diagnosis, data mining, clustering, classification.

INTRODUCTION

The paper is not intended to provide a comprehensive overview but rather describes some subareas and directions which from my personal point of view seem to be important for applying machine learning in medical diagnosis. Due to the difficulty of affinity propagation in dynamic data we use message passing concept compared with the previous related work to perform this in our project we can take the concept of clustering in message passing.

Clustering is an important concept in data mining, partitioning of data set into group are called clusters. Similar data are put into a group. There are different types of clustering and they are designed for discovering pattern in static data.

Classification is the predicting a certain outcome based on a given input. To predict the outcome, the algorithm processes a training set containing a set of attribute as a prediction attribute. Prediction rules can b expressed in the form of IF-THEN,

IF => Represent conjunction of condition.
THEN => Prediction attribute value for item that satisfy if condition.

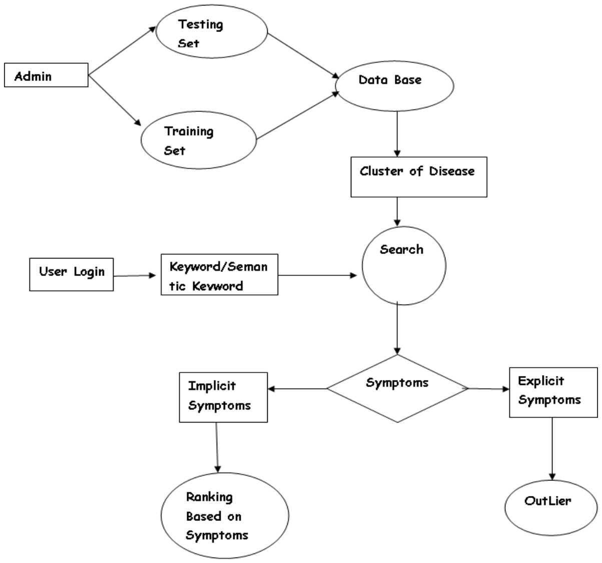
CONCEPT EXTRACTION

When a global online search is done, multiple results will be displayed and the relevant information that the user need can be in any of the page and there are chance to get the correct result even in any of the pages. To overcome this we use concept of outlier is used. In the project own search engine is created and when the user gives the symptoms in the search engine, the server will identify weather the user has the disease or not and it will produce the correct result.

For example consider the project based on cancer. There are many types of cancers and there are symptoms and cure for it. So when the user gives the symptoms in the search engine, the type of cancer will be displayed with the ranking. The cancer that is displayed will have the list of cancer with prevent, cure and side effects for the given symptoms. Outlier concept is used when the user give the symptoms with disease and symptoms with no disease.

Two sets will be created as training set and testing set. The training set contains cancer disease, symptoms and input for the machine language. The testing set has the patient information and the disease with prevent and cure. So when the users search with symptoms the disease will be displayed with ranking.

SYSTEM DESCRIPTION



The user search with the keyword with object, attribute, categories and clustering process is done and message passing is done and implementation of disease diagnosis process. Consider disease name, symptoms and biomedical analysis for automatic disease diagnosis process and in this the result are added which show the disease and symptom with prevent and cure.

RELATED WORK

Characteristics of the dynamic data requires is, the ability to adapt to changes in the data distribution. The ability to detect emerging clusters and distinguish them from outliers in the data. The ability to merge old clusters or discard expired ones. All of these requirements make dynamic data clustering a significant challenge.

In this existing system, there is no clustering techniques are followed in message passing. Problem of processing time stamp data, to produce a sequence of clustering result. The proposed model is based on the affinity propagation, where the newly arrived objects were clustered. In this each data sets were categorized as,

1. Varieties namely Categories (Product Name), 2. Objects (Variety-example Manufacturers), 3. Attributes (Sub Category - example Model Number).

Based on these three dataset clustering was formed, if a dataset is not fit into these three categories it will be considered as outlier and the data will not pass to the user.

ALGORITHM

Stemming is the term used in linguistic morphology and information retrieval to describe the process for reducing inflected (or sometimes derived) words to their word stem, base or root form generally a written word form. The stem needs not to be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root. Algorithms for stemming have been studied in computer science since the 1960s. Many search engines treat words with the same stem as synonyms as a kind of query expansion, a process called conflation.

Ranking algorithms With the growing number of Web pages and users on the Web, the number of queries submitted to the search engines are also growing rapidly day by day. Therefore, the search engines needs to be more efficient in its processing way and output. Web mining techniques are employed by the search engines to extract relevant documents from the web database documents and provide the necessary and required information to the users.

The search engines become very successful and popular if they use efficient ranking mechanisms. Now these days it is very successful because of its Page Rank algorithm. Page ranking algorithms are used by the search engines to present the search results by considering the relevance, importance and content score and web mining techniques to order them according to the user interest. Some ranking algorithms depend only on the link structure of the documents i.e. their popularity scores (web structure mining), whereas others look for the actual content in the documents (web content mining), while some use a combination of both i.e. they use content of the document as well as the link structure to assign a rank value for a given document.

If the search results are not displayed according to the user interest then the search engine will lose its popularity. So the ranking algorithms become very important.

$$PR(N) = \sum_{m \in Bn} PR(M)/L(M)$$

Where the page rank value for a web page u is dependent on the page rank values for each web page v out of the set Bn (This set contains all pages linking to web page N), divided by the number L (M) of links from page M.

The naive Bayesian classifier I limit the historical overview of statistical methods to the naive Bayesian classifier. From the very beginning I was very interested in it. The algorithm is extremely simple but very powerful, and later I discovered that it can provide also comprehensive explanations which were confirmed in long discussions with physicians. I was fascinated with its efficiency and ability to outperform most advanced and sophisticated algorithms in many medical and also non-medical diagnostic problems.

For example, when compared with six algorithms, described in the naive Bayesian classifier outperformed all the algorithms on five out of eight medical diagnostic problems (Kononenko *et al*., 1998). Another example is a hard problem in mechanical engineering, called mesh design. In one study, sophisticated inductive logic programming algorithms achieved modest classification accuracy between 12 and 29% (Lavarack and Dˇzeroski, 1994; Pompe and Kononenko, 1997) while the naive Bayesian classifier achieved 35%. The naive Bayesian classifier became for me a benchmark algorithm that in any medical domain has to be tried before any other advanced method.

Other researcher had similar experience. For example, Spiegel halter *et al*. (1993) were for several man-months developing an expert system based on Bayesian belief networks for diagnosing the heart disease for new-born babies. The final classification accuracy of the system was 65.5%. When they tried the naive Bayesian classifier, they obtained 67.3%. The theoretical basis for the successful applications of the naive Bayesian classifier (also called simple Bayes) and its variants was developed by Good (1950; 1964).

We demonstrated the efficiency of this approach in medical diagnosis and other applications (Kononenko *et al*., 1984; Cestnik *et al*., 1987). But only in the early nineties the issue of the transparency (in terms of the sum of information gains in favor or against a given decision) of this approach was also addressed and shown successful in the applications in medical diagnosis (Kononenko, 1989; 1993). This issue is addressed in more detail in Section 3.4 and illustrated in Table-2. Lately, various variants and extensions of the naive Bayesian classifier have been developed. Cestnik (1990) developed the m-estimate of probabilities that significantly improved the performance of the naive Bayesian classifiers in several medical problems. Kononenko (1991) developed a semi naive Bayesian classifier that goes beyond the “naivety” and detects dependencies between attributes. The advantage of fuzzy discretization of continuous attributes within the naive Bayesian classifier is described in (Kononenko, 1992). Langley (1993) developed a system that uses the naive Bayesian classifier in the nodes of the decision tree.

Pazzani (1997) developed another method for explicit searching of dependencies between attributes in the naive Bayesian classifier. The transparency of the naive Bayesian classifier can be further improved with the appropriate visualization (Kohavi *et al*., 1997).

A classifier that uses the naive Bayesian formula to calculate the probability of each class C given the values Vi of all the attributes for an instance to be classified, assuming the conditional independence of the attributes given the class,

$$P(C|V_1..V_n) = P(C)\prod_i \frac{P(C|V_i)}{P(C)}$$

PERFORMANCE

Performance testing to determine system performs in terms of responsiveness and stability under a particular workload. System performs well in web environment. Performance analysis approaches validates accuracy and efficiency of active learning leading to reliable and authentic predictions. Performance measures by f-measure, a measure that combines precision and recall ie fraction of retrieved documents and relevant documents. Precision takes all retrieved documents and recall takes data relevant to query that are successfully retrieved.

DEVELOPING AND VALIDATING CLASSIFIERS

Developing a classifier using SVMs or other classification technique consists of several steps: (a) choosing a method of analysis; (b) choosing a set of features or attributes that will be used to classify the subjects; (c) training the classifier; (d) validating the classifier; and (e) evaluating potential errors in the classification. Each step presents opportunities to introduce bias and error into the process.

EXPERIMENTAL OUTCOME

The Proposed system has analyzed the performance of our proposal against an outlier approach for genetic algorithm. The results obtained by the analyzed algorithms are shown in the Table-1, When an search is done in the search engine, the result will be displayed according to the symptom that the user has searched and if that symptoms match the type of cancer, all the result will be displayed with the prevention and cure for it with the ranking and if the symptoms don’t match it will display as no disease to the user.

Table-1. A brief description of the five data sets.

Data set	Number of objects	Number of attributes	Number of categories	Usage of data set
Iris	150	4	3	whole
Wine	178	13	3	whole
WDBC	569	30	2	whole
Car	1728	6	4	partly
Yeast	1484	8	10	partly

Table-2. Experimental details.

Dataset	Number of initial objects	Number of new arriving objects	Preference coefficient
Iris	100	10	0.015
Wine	128	10	0.015
WDBC	469	20	0.017
Car	210	10	0.015
Yeast	552	20	0.011

CONCLUSIONS

The proposition is based on such an idea that”if two objects are similar, they should not only be clustered into the same group, but also have the same statuses”. Both the two ideas are significant, and will be very helpful in dynamic clustering design. Incremental clustering is only a branch of dynamic data clustering. Some other problems such as how to determine the value of preference p, how to measure similarity between objects, and how to extract features from time series are also of great importance.

FUTURE ENHANCEMENT

The possible future work for this paper includes considering the symptoms with both the disease and best drug for the treatment of that disease. And the treatment can be verified. This project can be taken to the next level by doing in big data.

REFERENCE

J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd edition, Morgan Kaufmann, 2011. p. 444.

T.W. Liao, ”Clustering of Time Series Data: A Survey,” Pattern Recognition. vol. 38, no. 11, pp. 1857-1874, November. 2005.

A.K. Jain, ”Data Clustering: 50 Years Beyond K-means,” Pattern Recognition Letters. vol. 31, no. 8, pp. 651-666, June 2009.

S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O Callaghan,” Clustering Data Streams: Theory and Practice,” IEEE Trans. Knowledge and Data Eng. vol. 15, no. 3, pp. 515-528, May 2003.

J. Beringer and E. Hullermeier, ”Online Clustering of Parallel Data Streams,” Data and Knowledge Engineering. vol. 58, no. 2, pp. 180-204, August 2006.

A. Likas, N. Vlassis, and J.J. Verbeek, ”The Global k-means Clustering Algorithm,” Pattern Recognition. vol. 36, no. 2, pp. 451-461, February 2003.

A.M. Alonso, J.R. Berrendero, A. Hernandez, A. Justel,” Time Series Clustering based on Forecast Densities,” Computational Statistics and Data Analysis. vol. 51, no. 2, pp. 762-776, November 2006.

B.J. Frey and D. Dueck, ”Response to Comment on ’Clustering by Passing Messages Between Data Points’,” Science. vol. 319, no. 5864, pp. 726a-726d, February 2008.

B.J. Frey and D. Dueck, ”Clustering by Passing Messages between Data Points,” Science. vol. 315, no. 5814, pp. 972-976, February 2007.

J. Pearl, ”Fusion, Propagation, and Structuring in Belief Networks,” Artificial Intelligence. vol. 29, no. 3, pp. 241-288, 1986.

F.R. Kschischang, B.J. Frey, and H.A. Loeliger, ”Factor Graphs and the Sum-product Algorithm,” IEEE Trans. Information Theory. vol. 47, no. 2, pp. 498-519, February 2001.

J.S. Yedidia, W.T. Freeman, and Y. Weiss, ”Constructing Free- Energy Approximations and Generalized Belief Propagation Algorithms,” IEEE Trans. Information Theory. vol. 51, no. 7, pp. 2282-2312, July 2005.