www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume - 3 Issue - 8 August, 2014 Page No. 7691-7698

Data Mining Approaches for Diabetes using Feature selection

Thangaraju P[1], NancyBharathi G[2]
Department of Computer Applications,
Bishop Heber College (Autonomous),
Trichirappalli-620

Abstract :

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining is applied to find patterns to help in the important tasks of medical diagnosis and treatment. This project aims for mining the diabetes types with the help of feature selection techniques. The main interest is to identify research goals. Data mining has played an important role in diabetes research. Data mining would be a valuable asset for diabetes researchers because it can unearth hidden knowledge from a huge amount of diabetes-related data. The results could be used for both scientific research and real-life practice to improve the quality of health care diabetes patients. This article describes applications of data mining for the analysis of blood glucose and diabetes mellitus data.

The main purpose of this paper is to predict how people with different age groups are being affected by diabetes based on their life style activities and to find out factors responsible for the individual to be diabetic. The Best First Search Technique is implemented in this approach to fetch the data from the database. This approach is used to retrieve data in efficient manner and also in fraction of seconds without any delay. The greedy forward selection method also implemented in this approach to update the data in the database

Index Terms — Data Mining, Diabetes, feature selection, Best first, Greedy forward selection

1. Introduction :

Diabetes mellitus, or simply diabetes, is a set of related diseases in which the body cannot regulate the amount of sugar in the blood . It is a group of metabolic diseases in which a person has high blood sugar, either because the body does not produce enough insulin, or because cells do not respond to the insulin that is produced. This high blood sugar produces the classical symptoms of polyuria, polydipsia and polyphagia . There are three main types of diabetes mellitus . Type 1

diabetes mellitus results from the body's failure to produce insulin, and presently requires the person to inject insulin or wear an insulin pump. This form was previously referred to as "insulin-dependent diabetes mellitus" or "juvenile diabetes". Type 2 diabetes mellitus results from insulin resistance, a condition in which cells fail to use insulin properly, sometimes combined with an absolute insulin deficiency. This form was previously referred to as non insulin- dependent diabetes mellitus or "adult-onset diabetes". The third main form, gestational diabetes occurs when pregnant women without a previous diagnosis of diabetes develop a high blood glucose level. It may precede development of type 2 diabetes mellitus. As of 2000 it was estimated that 171 million people globally suffered from diabetes or 2.8% of the population. Type-2 diabetes is the most common type worldwide . Figures for the year 2007 show that the 5 countries with the largest amount of people diagnosed with diabetes were India (40.9 million), China (38.9 million), US (19.2 million), Russia (9.6 million), and Germany (7.4 million) . Data Mining refers to extracting or mining knowledge from large amounts of data. The aim of data mining is to make sense of large amounts of mostly unsupervised data, in some domain. the data. feature selection usually require that the classes be defined based on the data attribute values. They often describe these classes by looking at the characteristics of data already known to belong to class. Knowledge Discovery in Databases is the process of finding useful information and patterns in data which involves Selection, Pre-processing, Transformation, Data Mining and Evaluation.

1.1. Data Mining And Kdd Process

Data mining is a detailed process of analyzing large amounts of data and picking out the relevant information. It refers to extracting or mining knowledge from large amounts of data [2]. It is the fundamental stage inside the process of extraction of useful and comprehensible knowledge, previously unknown, from large quantities of data stored in different formats, with the objective of improving the decision of organizations where the data can be collected. However data mining and overall process is known as Knowledge Discovery from Databases (KDD) is an expensive process, especially in the stages of business objectives elicitation, data mining objectives elicitation, and data preparation.

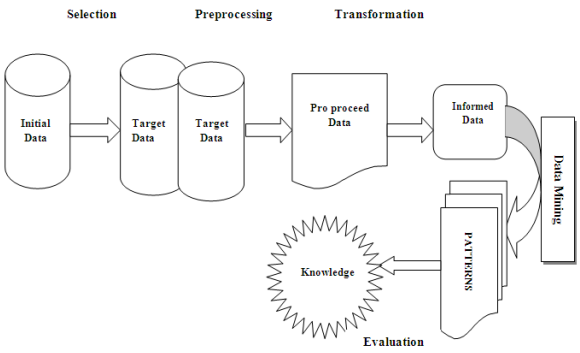


Figure 1.1 Data Mining is the core fo KDD Process

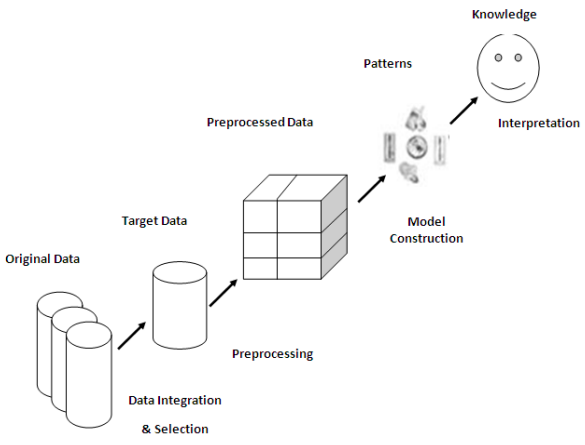


Figure 1.1a Processing and data mining

1.1 Classification

Classification is the process of categorization of data. Data can be classified according to any criteria, not only relative importance or frequency of use. Classification takes important role in data mining concept. There are many types of classification models are available. But the neural network and the decision tree are considered as widely used at the same time important models in classification.

1.2 Clustering:

Clustering is the process of grouping the related items. It is one of the useful and important type of technique for the process of discovery of data distribution. There are two major types of clustering.

1. Hierarchical clustering
2. Partitioning clustering

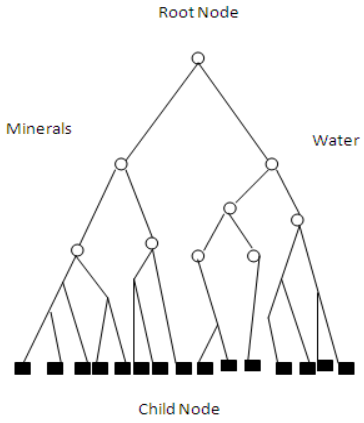


Figure 1.2 Hierarchical clustering

The process of partitioning the database into predefined numbers of clusters is simply called as hierarchical clustering. Hierarchical clustering do a sequence of partition is nestled into next partition in sequence.

1.3 Diabetes:

Diabetes Mellitus(DM) is also known as simply diabetes. It is a group of metabolic diseases in which there are high blood sugar levels over a prolonged period. This high blood sugar produces the symptoms of frequent urination increased thirst and increase hunger. Untreated diabetes can cause many complications. Acute complications include diabetic ketoacidosis and nonketotic hyperosmolar coma.

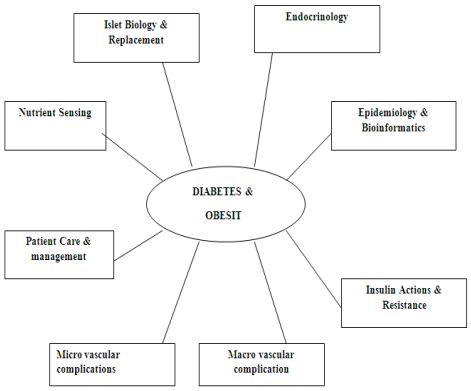


Figure 1.3 Diabetes

Feature selection

In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in model construction. The central assumption when using a feature selection technique is that the data contains many redundant or irrelevant features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context. Feature selection techniques are a subset of the more general field of feature extraction. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the features. Feature selection techniques are often used in domains where there are many features and comparatively few samples (or data points). The archetypal case is the use of feature selection in analysing DNA microarrays, where there are many thousands of features, and a few tens to hundreds of samples. Feature selection techniques provide three main benefits when constructing predictive models:

- improved model interpretability,
- shorter training times,

- enhanced generalisation by reducing overfitting.

Search-Based Techniques

search-based techniques are based on targeted projection pursuit which finds low-dimensional projections of the data that score highly: the features that have the largest projections in the lower-dimensional space are then selected.

1. Exhaustive
2. **Best first**
3. Simulated annealing
4. Genetic algorithm
5. **Greedy forward selection**
6. Greedy backward elimination
7. Targeted projection pursuit
8. Scatter Search[4]
9. Variable Neighborhood Search[5]

Best-first search

Best-first search is a search algorithm which explores a graph by expanding the most promising node chosen according to a specified rule.

Algorithm [3]

OPEN = [initial state]

while OPEN is not empty or until a goal is found do

1. Remove the best node from OPEN, call it n.
2. If n is the goal state, backtrace path to n (through recorded parents) and return path.
3. Create n's successors.
4. Evaluate each successor, add it to OPEN, and record its parent.

done

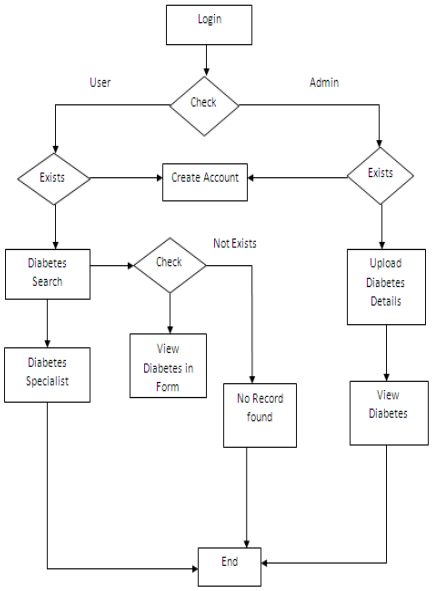


Figure1 Flow diagram

Feature selection is the process of selecting a subset of the terms occurring in the training set and using only this subset as features in text classification. Feature selection serves two main purposes. First, it makes training and applying a classifier more efficient by decreasing the size of the effective vocabulary. This is of particular importance for classifiers that, unlike NB, are expensive to train. Second, feature selection often increases classification accuracy by eliminating noise features. A noise feature is one that, when added to the document representation, increases the classification error on new data. Suppose a rare term, say arachnocentric, has no information about a class, say China, but all instances of arachnocentric happen to occur in China documents in our training set. Then the learning method might produce a classifier that misassigns test documents containing arachnocentric to China. Such an incorrect generalization from an

accidental property of the training set is called overfitting . We

SELECTFEATURES (D, c, k)

1. V ← EXTRACTVOCABULARY(D)
2. L ← [ ]
3. for each t ∈ V
4. do A(t, c) ← COMPUTEFEATUREUTILITY(D,c,k)
5. APPEND(L,(A(t,c),t))
6. return FEATURESWITHLARGESTVALUES(L,K)

Figure: Basic feature selection algorithm for selecting the k best features.

can view feature selection as a method for replacing a complex classifier (using all features) with a simpler one (using a subset of the features). It may appear counterintuitive at first that a seemingly weaker classifier is advantageous in statistical text classification, but when discussing the bias-variance tradeoff .we will see that weaker models are often preferable when limited training data are available.

The basic feature selection algorithm is shown in Figure1.0 . For a given class c, we compute a utility measure A(t,c)for each term of the vocabulary and select the k terms that have the highest values of A(t,c) All other terms are discarded and not used in classification. We will introduce three different utility measures in this section: mutual information, A (t, c) = I (U_t , C_c) ;X^2 the test, A (t, c) = X^2 (t, c); and frequency, A (t, c) = N (t, c) Of the two NB models, the Bernoulli model is particularly sensitive to noise features. A Bernoulli NB classifier requires some form of feature selection or else its accuracy will be low. This section mainly addresses feature selection for two-class classification tasks like China versus not-China. Section 1.0 briefly discusses optimizations for systems with more than two classes.

4.3 Results and Discussion

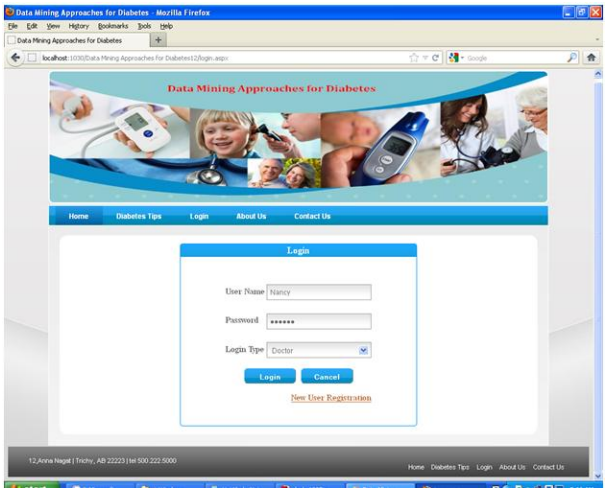


Figure 4.1 : Doctors and User Login

The Authentication module provides intact security control to our website. In order to verify that only authorized users are entered into our website, the authorization module will help us. The User Id and Password of the visiting user will be verified in this module, and on the positive result of the verification, the user will be allowed to browse further pages of our website. The new user has to register the prior information first. After registration they will get a user name and password.

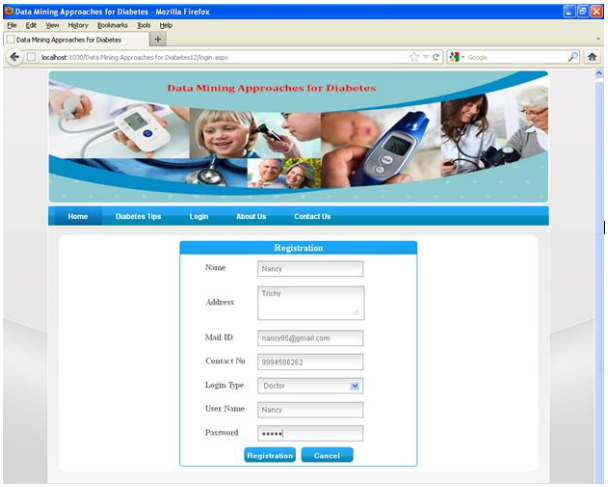
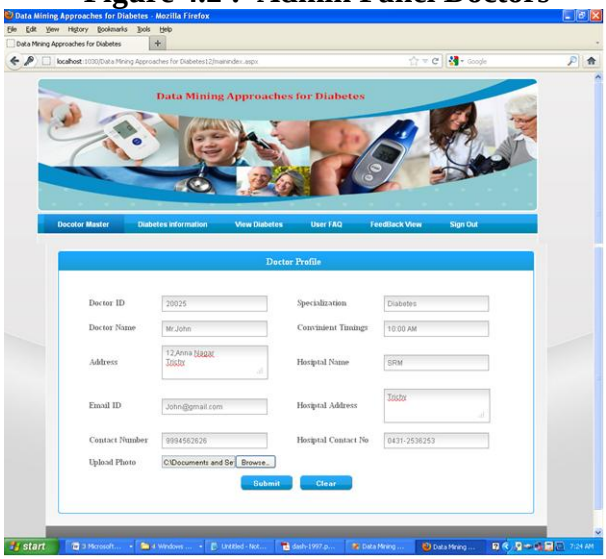


Figure 4.2 : Doctors and User Registration

This New User Registration module is used to register the details of the user. A new user has to register their details in this page. A user should enter all their required information without

fail. The administrator maintains all user information.

Figure 4.2 : Admin Panel Doctors



4.2 : Doctors Masters

In this module, Doctors, Medical experts want to upload the diseases details and treatment details. For that purpose they first registered his/her personal details and then login into page to upload medical files.

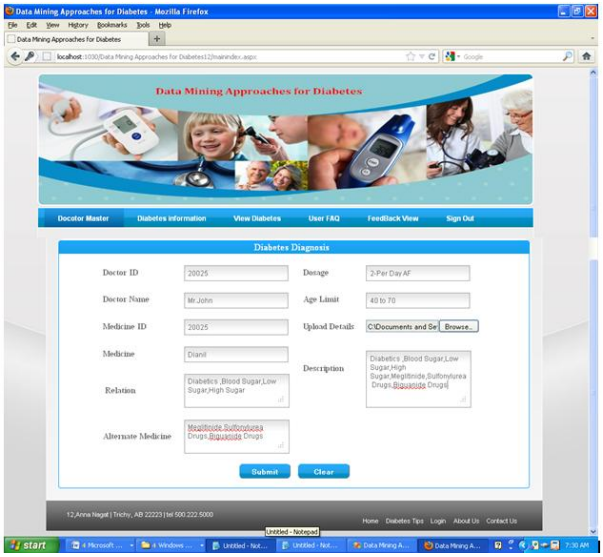


Figure 4.2 : Diabetes Masters

In this module, we describe about that the users are upload the medical details like disease details, with the treatment details, etc., it is very useful for the many people. They can easily retrieve the data's from the upload database. Each disease have separate id to generate. It stores the data in the database.

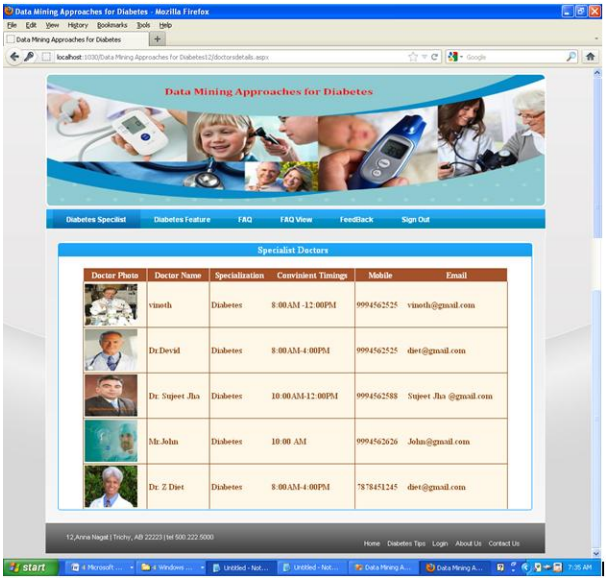


Figure 4.2 : Specialist Diabetes

Figure 4.2 Feature Selection Diabetes Information Search

In this module, the text relation fetcher is designed to search for information on the databases like World Wide Web. The search results are generally presented in a list of results often referred to as SERPS, or "search engine results pages". The information may consist of web pages, images, information and other types of files. Some search engines also mine data available in databases or open directories. Unlike web directories, which are maintained only by human editors, search engines also maintain real-time information by running an algorithm on a web crawler.

Conclusion

The conclusions of our study suggest that domain-specific knowledge improves the results. Probabilistic models are stable and reliable for tasks performed on short texts in the medical domain. The representation techniques influence the results of the feature selection Best Search algorithms, but more informative representations are the ones that consistently obtain the best results. The first task that we tackle in this paper is a task that has applications in information retrieval, information extraction, and text summarization. We identify potential improvements in results when more information is brought in the representation technique for the task of classifying short medical texts.

References

[1] Aha, D.W., Kibler, D. and Albert, M.K., Instance-Based Learning Algorithms. Machine Learning, 6:37–66, 1991.

[2] Bobrowski, L., Feature selection based on some homogeneity coefficient. In: Proceedings of Ninth International Conference on Pattern Recognition, 544–546, 1988.

[3] Brassard, G., and Bratley, P., Fundamentals of Algorithms. Prentice Hall, New Jersey, 1996.

4] Caruana, R. and Freitag, D., Greedy attribute selection.
In: Proceedings of Eleventh International Conference on Machine Learning, Morgan Kaufmann, New Brunswick**,** New Jersey, 28–36, 1994.

[5] Doak, J., An evaluation of feature selection methods and their application to computer security. Technical report, Davis,CA:

University of California, Department of Computer Science, 1992.

[6] Foroutan, I. and Sklansky, J., Feature selection for automatic classification of non-gaussian data. IEEE Transactions on Systems, Man, and Cybernatics, SMC 17(2):187–198, 1987.M. Dash, H. Liu / Intelligent Data Analysis 1 (1997) 131–156 155

[7] Ichino, M. and Sklansky, J., Feature selection for linear classifier. In: Proceedings of the Seventh International Conference on Pattern Recognition, volume 1, 124–127, July–Aug 1984.

[8] Kira, K. and Rendell, L.A., The feature selection problem: Traditional methods and a new algorithm. In: Proceedings of Ninth National Conference on Artificial Intelligence, 129–134, 1992.

[9] Liu, H. and Setiono, R., Feature selection and classification—a probabilistic wrapper approach. In: Proceedings of Ninth International Conference on Industrial and Engineering Applications of AI and ES, 284–292, 1996.

[10] Liu, H. and Setiono, R., A probabilistic approach to feature selection—a filter solution. In: Proceedings of International Conference on Machine Learning, 319–327, 1996.

[11] Modrzejewski, M., Feature selection using rough sets theory. In: Proceedings of the European Conference on Machine Learning (P. B. Brazdil, ed.), 213–226, 1993.

[12] Pagallo, G. and Haussler, D., Boolean feature discovery in empirical learning. Machine Learning, 1(1):81–106, 1986.

[13] Queiros, C.E. and Gelsema, E.S., On feature selection. In: Proceedings of Seventh International Conference on Pattern Recognition, 1:128–130, July-Aug 1984.

[14] Quinlan, J.R., C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, California, 1993.

[15] Segen, J., Feature selection and constructive inference. In: Proceedings of Seventh International Conference on Pattern Recognition, 1344–1346, 1984. 156 M. Dash, H. Liu / Intelligent Data Analysis 1 (1997) 131–156

[16] Sheinvald, J., Dom, B. and Niblack, W., A modelling approach to feature selection. In: Proceedings of Tenth International Conference on Pattern Recognition, 1:535–539, June 1990.

[17] Siedlecki, W. and Sklansky, J., On automatic feature selection. International Journal of Pattern Recognition and Artificial Intelligence, 2:197–220, 1988.

[18] Skalak, D.B., Prototype and feature selection by sampling and random mutation hill-climbing algorithms. In: Proceedings of Eleventh International Conference on Machine Learning, Morgan Kaufmann, New Brunswick, 293–301, 1994.

[19] Vafaie, H. and Imam, I.F., Feature selection methods: genetic algorithms vs. greedy-like search. In: Proceedings of International Conference on Fuzzy and Intelligent Control Systems, 1994.

[20] Xu, L., Yan, P. and Chang, T., Best first strategy for feature selection. In: Proceedings of Ninth International Conference on Pattern Recognition, 706–708, 1988.