ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 19, April 2016

IDENTIFICATION OF MEDICINE BEHAVIOUR WITH PREVENTION OF DISEASE RISK MANAGEMENT SYSTEM

[1]D. STERLIN RANI, [2] S.BENITTA SHERINE, [3] M.GEETHA PRIYA, [4] G. DIVYA
[1]ASSISTANT PROFESSOR, Department of CSE, KINGS ENGINEERING COLLEGE
[2, 3, 4] UG STUDENTS, Department of CSE, KINGS ENGINEERING COLLEGE

Abstract— **In the Early detection of patients with elevated risk of developing diabetes mellitus is critical to the improved prevention and overall clinical management of these patients. In the proposed system model We aim to apply association rule mining to electronic medical records (EMR) to discover sets of risk factors and their corresponding subpopulations that represent patients at particularly high risk of developing diabetes. In the modification process, Based on the affinity propagation the newly arrived objects were clustered. In this each data sets were categorized into three major Varieties namely Symptoms (Name), Drug (Variety – Eg. Manufacturers), Dosage and period (Number of day) &. Based on these three clustering was formed an dataset .Based on data set we analysis whether the it create the following diseases like BP, Cholesterol, Diabetes. And we also recommend medicine for the particular diseases. If these are not fit into these three categories it will be considered as outlier and the data will not pass to the user. We will also encrypt the Data for Ensuring Security.**

Index Terms—**Affinity propagation, Electronic medical records(EMR), Naive Bayes, Stemming.**

1. INTRODUCTION

Method of analyzing large numbers of Google search queries to track influenza-like illness in a population. This approach may make it possible to utilize search queries to detect influenza epidemics in areas with a large population of web search users. Extracting relations between entities from biomedical literature based solely on shallow linguistic information. We use a combination of kernel functions to integrate two different information sources: (i) the whole sentence where the relation appears, and (ii) the local contexts around the interacting entities. We performed experiments on extracting gene and protein interactions from two different data sets. The results show that our approach outperforms most of the previous methods based on syntactic and semantic information on the observation that the information required to assert a relationship between two named entities in the same sentence is typically captured by the shortest path between the two entities in the dependency graph. Entity relation detection is a form of information extraction that finds predefined relations between pairs of entities in text. This paper describes a relation detection approach that combines clues from different levels of syntactic processing using kernel methods. Information from three different levels of processing is considered: tokenization, sentence parsing and deep dependency analysis. Each source of information is represented by kernel functions. Then composite kernels are developed to integrate and extend individual kernels so that processing errors occurring at one level can be overcome by information from other levels. We present an evaluation of these methods on the 2004 ACE relation detection task, using Support Vector Machines, and show that each level of syntactic processing contributes useful information for this task. When evaluated on the official test data, our approach produced very competitive ACE value scores. We also compare the SVM with KNN on different kernels.

DIABETES mellitus is a growing epidemic that affects 25.8 million people in the U.S. (8% of the population), and approximately 7 million of them do not know they have the disease. Diabetes leads to significant medical complications including ischemic heart disease, stroke, nephropathy, retinopathy, neuropathy and peripheral vascular disease. Early identification of patients at risk of developing diabetes is a major healthcare need. Appropriate management of patients at risk with lifestyle changes and/or medications can decrease the risk of developing diabetes by 30% to 60%. Multiple risk factors have been identified affecting a large proportion of the population. For example, pre-diabetes (blood sugar levels above normal range but below the level of criteria for diabetes) is present in approximately 35% of the adult population and increases the absolute risk of diabetes 3 to 10 fold depending on the presence of additional associated risk factors, such as obesity, hypertension, hyper lipidemia, etc. Comprehensive medical management of this large portion of the population to prevent diabetes represents an unbearable burden to the healthcare system. In response to the pressing need to identify patients at high risk of diabetes early, numerous diabetes risk indices (risk scores) have been developed. These scores only provide a quantification of the risk, they are not suggestive of the

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 19, April 2016

factors that may have caused the elevation of the risk. Moreover, these scores utilize individual risk factors in an additive fashion without taking interactions among them into account. Diabetes is part of the metabolic syndrome, which is a constellation of diseases including hyper lipidemia (elevated triglyceride and low HDL levels), hypertension (high blood pressure) and central obesity (with body mass index exceeding 30 kg/m2). These diseases interact with each other, with cardiac and vascular diseases and thus understanding and modeling these interactions is important. Association rules are implications that associate a set of potentially interacting conditions (e.g. high BMI and the presence of hypertension diagnosis) with elevated risk. The use of association rules is particularly beneficial, because in addition to quantifying the diabetes risk, they also readily provide the physician with a "justification", namely the associated set of conditions. This set of conditions can be used to guide treatment towards a more personalized and targeted preventive care or diabetes management. While association rules themselves can be easily interpreted, the resulting rule sets can sometimes be very large, eroding the interpretability of the rule set as a whole. Especially, in this work, we consider a rich set of risk factors, namely co-morbid diseases, laboratory results, medications and demographic information that are commonly available in electronic medical record (EMR) systems. With such an extensive set of risk factors, the set of discovered rules grows combinatorial large, to a size that severely hinders interpretation. To overcome this challenge, we applied rule set summarization techniques to compress the original rule set into a more compact set that can be interpreted with ease.

2. RELATED WORKS

A diabetes index is in essence a predictive model that assigns a score to a patient based on his estimated risk of diabetes. They conducted an extensive survey of diabetes indices describing the risk factors and the modeling technique that these indices utilized. They found that most indices were additive in nature and none of the surveyed indices have taken interactions among the risk factors into account. While we are not aware of any new diabetes index published after the survey, a recent study focusing on the metabolic syndrome (of which diabetes is a component) represents a significant development. We used association rule mining to systematically explore co-occurrences of diagnosis codes. The resulting association rules do not constitute a diabetes index because the study does not designate a particular outcome of interest and they do not assess or predict the risk of diabetes in patients, but they discovered some significant associations between diagnosis codes. We have recently undertaken a diabetes study where we aimed to discover the relationships among diseases in the metabolic syndrome. We used the same cohort as this current study, however, we included only eight diagnosis codes and age as predictors. We discovered association rules involving some of these eight diagnosis codes, assessed the risk of diabetes that these rules confer on patients and presented the rules as a progression graph depicting how patients progress from a healthy state towards diabetes. We demonstrated that the approach found clinically meaningful association rules that are consistent with our medical expectation. With only eight predictor variables, the size of the discovered rule set was modest–13 significant rules– and consequently, interpretation was straightforward. Naturally, no rule-set summarization was necessary.

3. BACKGROUND

In our current work, we extend our previous study by incorporating a wide variety of predictors alongside the diagnosis codes. We use the same methodology of combining survival analysis and distributional association rule mining. In this section, we review the basic concepts underlying this methodology: Navie bayes classifier, stemming and we describe how these two techniques can be combined to find as rules implemented with survival outcomes.

3.1 NAIVE BAYES CLASSIFIER

Naïve Bayes: the conditional independence assumption Training is very easy and fast; just requiring considering each attribute in each class separately Test is straightforward; just looking up tables or calculating conditional probabilities with estimated distributions. A popular generative model. Performance competitive to most of state-of-the-art classifiers even in presence of violating independence assumption Many successful applications, e.g., spam mail filtering A good candidate of a base learner in ensemble learning Apart from classification, naïve Bayes can do more. MAP classification rule. MAP: Maximum A Posterior Assign x to c* if

$$P(C = c^* \mid \mathbf{X} = \mathbf{x}) > P(C = c \mid \mathbf{X} = \mathbf{x}) c \neq c^*, \; c = c$$

Generative classification with the MAP rule Apply Bayesian rule to convert them into posterior probabilities Then apply the MAP rule :

$$P(C = c_i \mid \mathbf{X} = \mathbf{x}) = \frac{P(\mathbf{X} = \mathbf{x} \mid C = c_i) P(C = c_i)}{P(\mathbf{X} = \mathbf{x})}$$
$$\propto P(\mathbf{X} = \mathbf{x} \mid C = c_i) P(C = c_i)$$
$$\text{for } i = 1,2,\cdots, L$$

Bayes classification Difficulty: learning the joint probability Naïve Bayes classification Assumption that all input features are conditionally independent!

$$P(X_1, X_2, \cdots, X_n \mid C) = P(X_1 \mid X_2, \cdots, X_n, C) P(X_2, \cdots, X_n \mid C)$$
$$= P(X_1 \mid C) P(X_2, \cdots, X_n \mid C)$$

MAP classification rule: for X=(x1, x2,….xn) Discrete valued features with a learning phase gives a training set S of F features and L classes. The output will be given as F*L conditional probabilistic generative models.

For each target value of c_i ($c_i = c_1$

$\hat{P}(C = c_i) \leftarrow$ estimate $P(C = c)$ with

For every feature value x_{ik} X_i ($j = 1, \cdots, F; k =$

$\hat{P}(X_j = x \mid C = c_i) \leftarrow$ estimate $P(X = x \mid C$

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 19, April 2016

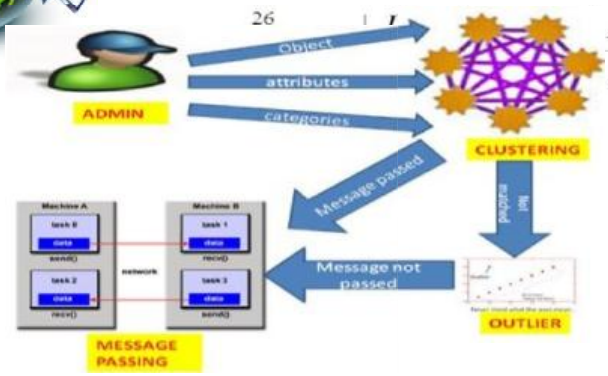

Fig: System Architecture

3.2 STEMMING

Stemming is one technique to provide ways of finding morphological variants of search terms. Used to improve retrieval effectiveness and to reduce the size of indexing files.

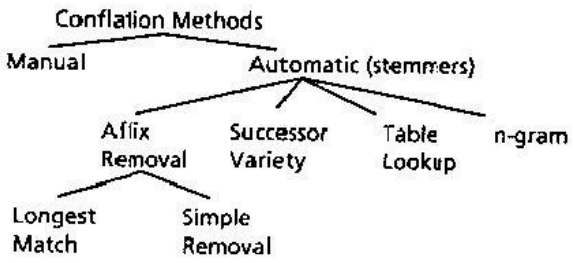

Fig: Taxonomy for stemming algorithms.

Criteria Overstemming: too much of a term is removed. Understemming: too little of a term is removed. Retrieval effectiveness measured with recall and precision, and on their speed, size, and so on *<X> ---the stem ends with a given letter X.*v*---the stem contains a vowel.*d ---the stem ends in double consonant. Suffix conditions take the form: (current_suffix == pattern). The majority of stemming's affection on retrieval performance have been positive. Stemming is as effective as manual conflation. The effect of stemming is dependent on the nature of vocabulary used. There appears to be little difference between the retrieval effectiveness of different full stemmers. Stemmers are used to conflate terms to improve retrieval effectiveness and /or to reduce the size of indexing file. Stemming will increase recall at the cost of decreased precision. Stemming can have marked effect on the size of indexing files, sometimes decreasing the size of file as much as 50 percent. Determine word and morpheme boundaries based on the distribution of phonemes in a large body of utterances. The successor variety of a string is the number of different characters that follow it in words in some body of text. The successor variety of substrings of a term will decrease as more characters are added until a segment boundary is reached. Entropy method - the number of words in a text body beginning with the i length sequence of letters. The number of words in with the successor j. The probability that a member of number of words in has the successor j is given by

$$\frac{|D\alpha ij|}{|D\alpha i|}$$

The entropy of $|D\alpha|$ is

$$\frac{|D\alpha ij|}{|D\alpha i|}$$

Two criteria used to evaluate various segmentation methods: the number of correct segment cuts divided by the total number of cuts. The number of correct segment cuts divided by the total number of true boundaries. After segmenting, if the first segment occurs in more than 12 words in the corpus, it is probably a prefix. The successor variety stemming process has three parts:

1. Determine the successor varieties for a word
2. Segment the word using one of the methods
3. Select one of the segments as the stem

4. CONCLUSION

The electronic data generated by the use of EMRs in routine clinical practice has the potential to facilitate the discovery of new knowledge. Association rule mining coupled to a summarization technique provides a critical tool for clinical research. It can uncover hidden clinical relationships and can propose new patterns of conditions to redirect prevention, management, and treatment approaches. For this method to be useful, the number of rules needed to be reduced to a level where clinical interpretation is feasible. To this end, we studied four methods to summarize these rules into sets of 10-20 rules that clinical investigators can evaluate. While all four methods created reasonable summaries, each method had its clear strength. However, not all of these strengths are necessarily beneficial to our application. We found that the most important differentiator between the algorithms is whether they use a selection criterion to include a rule in the summary based on the expression of the rule or based on the patient subpopulation that the rule covers.

REFERENCES

[1] F. Afrati, A. Gionis, and H. Mannila, "Approximating a collection of frequent sets," in Proc. ACM Int. Conf. KDD, Washington, DC, USA, 2004.

[2] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in Proc. 20th VLDB, Santiago, Chile, 1994.

[3] Y. Aumann and Y. Lindell, "A statistical theory for quantitative association rules," in Proc. 5th KDD, New York, NY, USA, 1999.

[4] P. J. Caraballo, M. R. Castro, S. S. Cha, P. W. Li, and G. J. Simon, "Use of association rule mining to assess diabetes risk in patients with impared fasting glucose," in Proc.

[5] V. Chandola and V. Kumar, "Summarization – Compressing data into an informative representation," Knowl. Inform. Syst., vol. 12, no. 3, pp. 355–378, 2006.

[6] G. S Collins, S. Mallett, O. Omar, and L.-M. Yu, "Developing risk prediction models for type 2 diabetes: A systematic review of methodology and reporting," BMC Med., 9:103, Sept. 2011.

[7] Diabetes Prevention Program Research Group, "Reduction indiabetes with lifestyle intervention ," N. Engl. J. Med., vol. 346, no. 6, pp. 393–403, Feb. 2002.