Improving a Novel Data Mining Approach Leveraging Social Media to Monitor Consumer Opinion of Sitagliptin

Sandeep Kaur¹ & Lalit Mann Singh²

¹School of Engineering, Sri Guru Granth Sahib World University, Fatehgarh Sahib.
²Assistant Professor, Department Of Computer Science and Engineering, Sri Guru Granth Sahib World University

Abstract: A novel statistics mining technique changed into evolved to gauge the enjoy of the drug sitagliptin (exchange call Januvia) through patients with diabetes mellitus kind 2. To this purpose, we devised a two-step evaluation framework. Initial exploratory evaluation using self-organizing maps changed into executed to decide structures primarily based on user reviews a number of the forum posts. The consequences have been a compilation of user’s clusters and their correlated (fantastic or bad) opinion of the drug. Subsequent modeling the use of community evaluation methods changed into used to decide influential customers among the discussion board members. These findings can open new avenues of studies into speedy facts collection, remarks, and evaluation that can allow advanced consequences and answers for public health and vital feedback for the manufacturer.

INTRODUCTION

Data mining is becoming more and more popular thanks to the rapid development of computers and the need to extract information out of increasingly large data collections. Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouse.

1.1 DIABETES MELLITUS

Diabetes mellitus is a group of metabolic disorders characterized by chronic hyperglycemia with disturbances of carbohydrate, fat and protein metabolism resulting from defects in insulin secretion, insulin action, or both (WHO 1999; American Diabetes Association 2008). Several pathogenic processes are involved in the development of diabetes. These include destruction of the beta cells of the pancreas with consequent insulin deficiency and abnormalities that result in resistance to insulin action.

1.1.1 Classification

Diabetes mellitus can be classified into four types based on etiology: type 1 diabetes, type 2 diabetes, gestational diabetes and other types of diabetes. The majority of cases of diabetes fall into type 1 diabetes or type 2 diabetes (American Diabetes Association 2008).

1.1.1.1 Type 1 diabetes

Type 1 diabetes is characterized by an absolute deficiency of insulin secretion (Notkins et al. 2001; American Diabetes Association 2008). The most common form of type 1 diabetes is immune-mediated diabetes and accounts for 5-10% of all individuals with diabetes. This disease is considered autoimmune (Rose et al. 1993; Bach 1994) and results from a cellular mediated autoimmune destruction of the beta cells of the pancreas (Knip 1997; Knip et al. 2008). Autoimmune destruction of beta cells has multiple genetic predispositions, mainly strong HLA associations, and is influenced by environmental factors (Knip 1997; Akerblom et al. 2002; Ilonen et al. 2002; Achenbach et al. 2005). Auto antibodies against beta-cell proteins, such as glutamic acid decarboxylase (GAD) (Baekkeskov et al. 1990), the tyrosine phosphates like protein islet antigen-2 (IA-2) (Lan et al. 1996; Notkins et al. 1996), insulin (Palmer et al. 1983) and zinc transporter 8 (ZnT8) (Wenzlau et al. 2007), are produced during the autoimmune destruction of the beta cells. One or more of these auto antibodies are present in 90-95% of individuals when hyperglycemia is initially detected (Notkins et al. 2001). Type 1 diabetes occurs predominantly in children and adolescents, usually as a rapidly progressive form, but may also occur in adults, often as a slowly progressive form, referred to as latent autoimmune diabetes in adults (LADA). A minority of patients with type 1 diabetes fall into the category named idiopathic...
diabetes (American Diabetes Association 2008). This form of diabetes is most common in non-
Caucasians and is strongly inherited, lacks immunological evidence for beta-cell autoimmunity, and is not HLA associated.

1.1.1.2 Type 2 diabetes
In type 2 diabetes, the cause is a combination of resistance to insulin action and defects in insulin secretion (American Diabetes Association 2008). In adults, type 2 diabetes is much more prevalent than type 1 diabetes, but it is rare in children and adolescents in Sweden. Individuals with type 2 diabetes do initially not need insulin treatment to survive, and often not later in life either. Most patients are obese, which causes some degree of insulin resistance. Type 2 diabetes is often associated with a strong genetic predisposition and the risk of developing this disease increases with age, obesity and lack of physical activity. Individuals with this form of diabetes are often undiagnosed for many years because their hyperglycemia is not severe enough to present symptoms, but they have increased risk to develop macro vascular and micro vascular complications. Insulin sensitivity may be increased, but not restored to normal, by weight reduction, increased physical activity and/or pharmacological treatment of hyperglycemia.

1.2 SELF ORGANIZING MAPs (SOM)

A SOM consists of neurons organized on a regular low dimensional grid, see Figure 1. Each neuron is a dimensional weight vector (prototype vector, codebook vector) where d is equal to the dimension of the input vectors. The neurons are connected to adjacent neurons by a neighborhood relation, which dictates the topology, or structure, of the map. In the Toolbox, topology is divided to two factors: local lattice structure (hexagonal or rectangular and global map shape (sheet, cylinder). The Self-Organizing Map (SOM) has been introduced in [Kohonen 1982] since then it has been analyzed and employed extensively. A recent overview can be found in [Kohonen 1997].

1.3 THE GENERATIVE TOPOGRAPHIC MAPPING (GTM)

Many data sets exhibit significant correlations between the variables. One way to capture such structure is to model the distribution of the data in terms of latent, or hidden, variables. A familiar example of this approach is factor analysis, which is based on a linear transformation from latent space to data space. In this paper we show how the latent variable framework can be extended to allow non-linear transformations while remaining computationally tractable. This leads to the GTM (Generative Topographic Mapping) algorithm, which is based on a constrained mixture of Gaussians whose parameters can be optimized using the EM (expectation-maximization) algorithm.

1.4 GENETIC ALGORITHM(GA)

There are no known polynomial time algorithms to solve many real world optimization problems making them hard to solve. A number of heuristics have been designed to solve the hard problems. These heuristics may provide sub optimal but acceptable solutions in a reasonable computational time. A number of Meta heuristics such as simulates annealing, evolutionary algorithms; artificial neural networks derived from natural physical and biological phenomena have also been used to solve these problems.

2. METHODS

2.1 FETCH THE REAL TIME DATA

The first step turned into to find the maximum popular discussion board committed to diabetes mellitus kind 2. We in comparison the quest of four maximum famous diabetes-associated message forums: DiabetesForum.com, Healthboards.com, Foum.lowcarber.org, and DiabetesDaily.com. A list of drugs and devices used by patients changed into compiled and searched in the outcomes of every of the message boards: the purpose become to examine which tablets and/or devices the sufferers were discussing the maximum. Sitagliptin became observed to be the most mentioned drug primarily based on the large variety of posts on the drug on the message boards. The message board DiabetesDaily.com changed into selected due to the fact Sitagliptin changed into the most regularly mentioned drug compared to the alternative four message forums. Within the message board DiabetesDaily.com, a seek using the exchange term ‘Januvia’ garnered more results (2600 consequences) than the drug time period ‘Sitagliptin’ (ninety two outcomes). The quested result the usage of ‘Januvia’ additionally again 713 posts, ranging from 2007 to the present time.

2.2 PREPROCESSING

After compiling the posts, they fed the list to reduce the size of data set by using the genetic algorithm. The whole data is divided into two parts training set and test set data.

2.3 APPLY GENETIC ALGORITHM

Genetic algorithms are inspired by Darwin’s theory about evolution. Solution to a problem solved by genetic algorithms is evolved. There are no known polynomial time algorithms to
solve many real world optimization problems making them hard to solve. A number of heuristics have been designed to solve the hard problems. These heuristics may provide sub optimal but acceptable solutions in a reasonable computational time. A number of Meta heuristics such as simulates annealing, evolutionary algorithms; artificial neural networks derived from natural physical and biological phenomena have also been used to solve these problems. GA is adaptive procedures derived from Darwin’s principal of survival of the fittest in natural genetics. GA maintains a population of potential solutions of the candidate problem termed as individuals. By manipulation of these individuals through genetic operators such as selection, crossover and mutation, GA evolves towards better solutions over a number of generations Algorithm is started with a set of solutions (represented by chromosomes) called population. Solutions from one population are taken and used to form a new population. This is motivated by a hope, that the new population will be better than the old one. Solutions which are selected to form new solutions (offspring) are selected according to their fitness - the more suitable they are the more chances they have to reproduce. This is repeated until some condition (for example number of populations or improvement of the best solution) is satisfied.

2.4 The Generative Topographic Mapping (GTM)

Many data sets exhibit significant correlations between the variables. One way to capture such structure is to model the distribution of the data in terms of latent, or hidden, variables. A familiar example of this approach is factor analysis, which is based on a linear transformation from latent space to data space. In this paper we show how the latent variable framework can be extended to allow non-linear transformations while remaining computationally tractable. This leads to the GTM (Generative Topographic Mapping) algorithm, which is based on a constrained mixture of Gaussians whose parameters can be optimized using the EM (expectation-maximization) algorithm.

One of the motivations for this work is to provide a principled alternative to the widely used ‘self organizing map’ (SOM) algorithm (Kohonen 1982) in which a set of unlabelled data vectors \( t_n \) (\( n = 1; \ldots ;N \)) in a D-dimensional data space is summarized in terms of a set of reference vectors having a spatial organization corresponding to a (generally) two-dimensional sheet. While this algorithm has achieved many successes in practical applications, it also suffers from some significant deficiencies, many of which are highlighted in Kohonen (1995). They include: the absence of a cost function, the lack of a theoretical basis for choosing learning rate parameter schedules and neighborhood parameters to ensure topographic ordering, the absence of any general proofs of convergence, and the fact that the model does not define a probability density. These problems can all be traced to the heuristic origins of the SOM algorithm. We show that the GTM algorithm overcomes most of the limitations of the SOM while introducing no significant disadvantages.

An important application of latent variable models is to data visualization. Many of the models used in visualization are regarded as defining a projection from the D-dimensional data space onto a two-dimensional visualization space. We shall see that, by contrast, the GTM model is defined in terms of a mapping from the latent space into the data space. For the purposes of data visualization, the mapping is then inverted using Bayes’ theorem, giving rise to a posterior distribution in latent space.

2.5 RESULT ANALYSIS

GTM offers more information about the mapping than the SOM because it calculates the distribution of single data points. The SOM can be seen as an approximation of the GTM. However it is not easy to visualize this additional information and often leads to simplifications in which this information is lost. One advantage GTM offers due to its statistical background is that the posterior probability of the GTM can be easily used to compare different models. Following table show that proposed system is better than existing one in term of accuracy, precision and recall.

3. LITERATURE SURVEY

A. Akay et al.[4] They devised a two-step evaluation framework. Preliminary exploratory evaluation using self-organizing maps became achieved to determine systems based on user evaluations some of the forum posts. The consequences had been a compilation of person’s clusters and their correlated (positive or terrible) opinion of the drug. Next modeling the usage of community analysis strategies became used to decide influential users a number of the forum participants. Those findings can open new avenues of research into fast statistics collection, remarks, and analysis which could permit stepped forward outcomes and answers for public fitness and crucial comments for the producer Zoe Lacroix et al.[5]. They model the records items in those resources and the hyperlinks between objects as an object graph. They become aware of a set of thrilling houses for hyperlinks and paths, including out degree, photo of a hyperlink, cardinality of statistics gadgets and hyperlinks, the
variety of wonderful items reached through some links, etc. Analogous to database cost fashions, we use information from the item graph to expand a framework to estimate the end result length for a query at the item graph. Analogous to schooling and testing, we use sampled facts from queries to estimate the end result length. They validate our models using data sampled from 4 NIH/NCBI information sources. Their study presents a foundation for querying and exploring statistics assets.

Noesnser, Jan, et al [7]. They endorse a singular technique to item reconciliation that is primarily based on an existing semantic similarity degree for connected statistics. They adapt the degree to the item reconciliation hassle, present precise and approximate algorithms that successfully implement the techniques, and provide a scientific experimental evaluation based on a benchmark dataset. As their predominant end result, they show that using lightweight ontology’s and schema facts substantially improves object reconciliation within the context of connected open statistics.

LiseGetoor, Christopher P. Diehl [68]. Numerous datasets of interest today are best portrayed as a connected gathering of interrelated items. These might speak to homogeneous systems, in which there is a solitary article sort and connection sort, or wealthier, heterogeneous systems, in which there might be different question and connection sorts (and conceivably other semantic data). Illustrations of homogeneous systems incorporate single mode informal organizations, such as individuals associated by kinship joins, or the WWW, a gathering of connected site pages. Illustrations of heterogeneous systems incorporate those in restorative areas portraying patients, ailments, medicines and contacts, or in bibliographic areas portraying productions, creators, and venues. Join mining alludes to information mining procedures that expressly consider these connections when building prescient or clear models of the connected information. Generally tended to connection mining errands incorporate article positioning, bunch recognition, aggregate classification, join forecast and sub graph revelation. While system investigation has been concentrated on top to bottom specifically territories for example, interpersonal organization examination, hypertext mining, and web investigation, just as of late has there been a cross-preparation of thoughts among these different groups. This is an energizing, quickly extending region. In this article, we survey a portion of the normal rising subjects.

Noesnser, Jan, et al [7]. They increase the analysis and display the way it gives insight into approaches of designing solid hyperlink analysis techniques. This in flip motivates new algorithms, whose overall performance they examine empirically the usage of citation facts and web hyperlink statistics Holmberg, H., Vaara, O.[67]. concluded that on the evidence available in 2001, it was not clear whether screening would be worthwhile. This paper summaries a review done for the UK NSC, and is structured along the lines of the NSC criteria. They screening for T2DM meets many of the criteria, but that number of problem remain.

Engelgau and Colleagues[34]. Carrier out a series of reviews of screening for diabetes. In the version of 2000, they provide a thorough review examining how well screening criteria. Although this was balanced are well argued review, it cannot be classed as systematic; no detail of search strategy of inclusion and exclusion criteria. Harris and Colleagues[9]. They provided a high quality systematic review, with details on sources searched and search strategy, study selection, data extraction and synthesis and other aspects of methods.

4 CONCLUSIONS

GTM gives a scattered representation of the positive and negative words group in the forum Diabetes Daily. Prior to the final GTM, a subset of the data was used for training the GTM. This was to ensure that GTM was trained to accurately model a sample set of the data prior to receiving the whole data set. In end, the data were selected for training the GTM. We used a 13 x 11 map size with four hundred variables from the modified wordlist to ascertain the weight of the words corresponded to the opinion of the drug Sitagliptin. A criterion for selecting the variables was that each word should appear ten times and above. This allowed us to achieve a uniform set of measurements while eliminating statistically insignificant outliers. The bulk of the user’s posts converged on four points of the map. We checked the correlation of the respective nodes with the values of their weight vectors corresponding to positive or negative words. This is how we defined the positive and negative areas of the map. A picture begins to emerge of user opinion that is roughly divided with regards to satisfaction of the drug Sitagliptin. One source of negative opinion stems from the side effects of the drug. A review of the medical literature has confirmed the very same side effects that the users were discussing [37-40]. Other sources of negative opinion vary from user frustration of the drug costs to frustration at the medical community. Positive opinions mainly stemmed from satisfaction by users who switched to it based on recommendations from a physician.
Figure 1: Represent the result of SOM.

Figure 2: Represent the result of GTM.

It shows that GTM offers more information about the mapping than the SOM because it calculates the distribution of single data points. The SOM can be seen as an approximation of the GTM. However it is not easy to visualize this additional information and often leads to simplifications in which this information is lost. One advantage GTM offers due to its statistical background is that the posterior probability of the GTM can be easily used to compare different models. Following table show that proposed system is better than existing one in term of accuracy, precision and recall.

REFERENCES


[18] S.R. Das and M.Y. Chen, “Yahoo! for Amazon: Sentiment extraction from small talk on