International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-2S, December 2018

Effective Identification of Features for the Diagnosis of Parkinson's disease using High utility Item set Mining together with GMM

B. Mouleswararao, Y. Srinivas

Abstract*: Disease detection is an imperative task in medical discipline. Many techniques based on image processing and data mining were employed for the early disease detection. In recent years, in spite of the latest encroachments in the science and technology, individuals experience from abundant brain disorders diseases such as Alzheimer and Parkinson. Among these diseases, Parkinson's Disease(P.D.) is mostly influenced around the world and therefore many methodologies were emerged to combat the disease. However, as the number of symptoms prevailing to this disease is plentiful, identifying the most subjective symptom is a challenging task. This article makes an attempt to identify the most prevailing symptoms based on high utility mining together with statistical modeling, such that effective treatment can be imparted at the early stage.

Index Terms: High utility item set, statistical modeling, Parkinson's disease, Alzheimer disease, medical imaging.

1. INTRODUCTION

World Health Organization has presented the most recent statistics about the most dangerous diseases subjected to the mankind. As per this survey, the diseases pertaining to nervous system ailments are identified to be dominating. These diseases persuade the mind and the vertebrae and identified that the spread of this disease has been radically amplified worldwide and roughly 14 out of 50 persons are suffering with neurological diseases [1]. A good number of these neurological chaoses are owed to the sway of P.D .and Alzheimer disease. P.D is a degenerative disease of the innermost nervous system associated with unrelenting and steady disorder in the movement of muscles. Various cases with P.D are observed worldwide and in the majority of cases, the distressed patient's age group is just about 50 years. The base is not exactly known and there is scarcely any treatment, nevertheless, premature diagnosis of the disease assist in the therapeutic and the treatment is supposed to be carried out till the patient's life time.

Many thoughts were projected in the literature using diverse theories like Artificial Intelligence[2],Particle Swam Algorithms [3], Data mining Algorithms[4], Statistical models [5] etc. Nevertheless, the instigators while suggesting their thoughts have taken various features pertaining to the sign into concern and allied with the disease [6]. Most of the offered works have taken in account lone features [7] and the works considered based on Apriori algorithms intended at classifying the most precise symptoms amid the a choice of symptoms linked with the disease, and this approach helped to resolve the concern to certain degree. However aprioiri algorithms experience some restrictions, such as, very lethargic computation, massive generation of subsets is an additional restraint [8]. To defeat this challenge FP-Growth algorithms have been initiated. FP-Growth algorithms can override the restrictions of Apriori algorithm and be capable to extract the relations without producing the candidate set. The key restrictions of the model is that it is extremely tricky to execute because it formulate the multifaceted data formation and takes enormous computational time to produce the FP-Tree and also needs hefty storage space for hoarding the data[9]. Therefore it is obligatory to build up well-organized methods that can help to spot the texture with alleviate. In this article we have considered high utility item set mining technique for the classification of the most influential indications that has foremost influence in discovery of the disease. These features are given to the Gaussian mixture model. The outcomes resulting after carrying out tests has helped to spot the main noteworthy symptoms. The manuscript is ordered as follows. In Section 2 of the document the notion of High utility item set mining(HUIM) are presented, Section 3 the data set of patients considered is presented. Section 4 of the document deal with the most customary symptoms, in Section 5, the experimentation is highlighted .The closing section 6, sums up the article.

2. HIGH UTILITY ITEM SET MINING

The concepts of data mining help to mine the patterns of interest and also help to group and classify the most significant patterns of voluminous datasets. The discovery of associations or the correlation between the various elements in the database is known as association rule mining, and finding the frequent element sets of these rules helps to discover the association rules in the databases. Several common element set mining algorithms are proposed to find most of the sets of repetitive elements in large databases. The FP growth algorithm uses a tree-based approach and scans the database in a deep way and generates a tree called FP-tree, which is an extended tree-prefix structure for storing frequent patterns. The property followed all common elements sets (FIM) mining algorithms to reduce the search space.

Manuscript published on 30 December 2018.
***** Correspondence Author (s)
B. Mouleswararao, Research Scholar, Department of CSE, GITAM University, Visakhapatnam, AP, India. (E-mail: mbpalli@gmail.com)
Y. Srinivas Professor, Department of IT, GITAM University, Visakhapatnam, AP, India (E-mail: sriteja.y@gmail.com)

FIM algorithms only generate sets of frequent elements, in which they consider the presence or absence of a particular element in a transaction. FIM algorithms also assume that each item appears in the transaction only once and all items will have same importance or weight or unit benefit or value, etc. They do not consider the fact that an article can appear more than once in a transaction and each article has an associated cost.For example, if a customer purchases two identical items in a single transaction and that item is purchased at a price, the FIM extraction algorithms does not consider this information as the frequency of an item in a single transaction and the cost of the item. This can lead to losing valuable information to discover sets of frequent items that generate a high profit. Therefore, you discover that items that are gaining high profits can not be discovered by using frequent article sets mining algorithms. To fill this gap, the problem of FIM has extended to the problem of Mining High Value Element Sets (HUIM) that considers the frequency of items that may appear more than once in a transaction and the cost of each element.

For each item in the transaction we consider two properties; cost /profit/importance/weight of item known as external utility and the frequency of the item in transaction, called as internal utility. The utility of an item can be calculated as the product of as external utility and the internal utility. The collection of items is known as item set. An item set is called a high utility item set if its utility is more than equals to a user given threshold. Otherwise, it is called as low utility item set. Finding high utility item sets from large databases is called High-Utility Item set Mining (HUIM). HUIM is widely used in areas like we click streaming analysis, cross-marketing in retail stores and bio-medical and medical applications.

In FIM, the close-down property is used to eliminate the search space when frequent item sets are found. High utility element sets can have supersets or subsets with a lower, equal or higher utility. Because the same downlink property can not be used in FIM, as in HUIM and the problem of finding sets of high utility elements is considered a more difficult problem compared to FIM. When the search space is very large and the database contains large transactions or a minimum threshold specified by the user is set to a very low value, the search for high utility element sets will become a Herculean task. Therefore, to reduce the search space effectively and to capture all sets of high-utility elements efficiently by not losing the vital information is a great challenge in mining sets of highly useful items.

Many investigations have been carried out for effective pruning in mining problems of sets of highly useful elements. The transaction-weighted closing property (TWDC) is proposed to eliminate the search space. TWDC property which states that, for any set of X elements, if X is not a set of weighted utility elements of high transactions, then any superset of X will not be a set of high utility items . The transaction-weighted utility (TWU) of a set of items is specified as the sum of the transaction utility values of all transactions in which the set of X items appears. TWDC will function as a down-lock property that we have been using in FIM to delete the search spaces. Many algorithms use the TWDC property to find sets of highly useful elements. In the present article, HUIM concepts are considered to identify the most appropriate symptoms associated with P.D.

2.1 Basic Definitions of High Utility Item set mining

First we define the utility mining problem related to P.D. and then describe the formal definitions specified in the literature as follows.

Let I = {i1, i2, ..., im} be a finite set of symptoms I, $1 \leq j \leq k$, and k is the length of X. A list with the length k is called k-joint symptom.∈Each symptom ik ($1 \leq k \leq m$) has a unit of importance p (ik) that is called External Utility of that symptom. A data set X is a set of k different symptoms {i1, i2, ..., ik} A transaction database D = {T1, T2, ..., Tn} contains a set of n patients, where each patient Tk ($1 \leq k \leq n$) is identified with a unique identifier Tk, called TID. Each ik symptom associated with the patient Tk is associated with a disease q (ik, Tk), that is, the frequency of having ik symptoms among Tk patients, which is also considered the most important reason for the onset of ik symptoms.

Table-1 : Example Patients Database D (Patients with Specific Diseases)

TID	TRANSACTION	Transaction Utility(TU)
T_1	(i_1, 2), (i_3,5), (i_4,3)	21
T_2	(i_1,3), (i_3,5), (i_5,1) (i_7,4)	27
T_3	(i_1,3), (i_2,4), (i_4,5) (i_5,3), (i_6,4)	50
T_4	(i_2,5), (i_3,4), (i_4,5), (i_5,5)	39
T_5	(i_2,3), (i_3, 3), (i_5,2), (i_7,3)	18
T_6	(i_1,2), (i_2,4), (i_4,3), (i_8,2)	26

Table-2: Most specific symptoms

Item	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
Profit	5	2	1	2	3	5	1	1

Definition1: The patient with a symptom i_p in the patient's dataset T_d is denoted as $u(i_p,T_d)$ and defined as $p(i_p) \times q(i_p,T_d)$.

Example 1:

Symptoms of each patient i_1 in transaction T_1= $q(i_1,T_1)=2$

Specific symptoms associated with each patient i_1 in transaction T_1= $p(i_1)=5$

Possibility of the disease with the symptoms for a patient i_1 in

$T_1 = u(i_1,T_1) = p(i_1) \times q(i_1,T_1) = 5 \times 2 = 10.$

Definition 2: The Most specific symptoms in the entire data set of patients, X in T_d, specified as as $u(X, T_d)$ and defined as

$$u(X, T_d) = \sum_{i_p \in X \wedge X \subseteq T_d} u(i, T_d)$$

Example2:

$u(\{i_1, i_4\},T_1)= u(\{i_1\},T_1) + u(\{i_4\},T_1) = 10+6 = 16.$

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Definition3: A patient is said to be disease prone, if he has a particular symptom X in D and is denoted as u(X) and defined as

$$u(X)= \sum_{X \subseteq T_d \wedge T_d \in D} u(i_p, T_d)$$

Example 3:

$u(\{i_1, i_4\})= u(\{i_1, i_4\},T_1) + u(\{i_1, i_4\},T_3) + u(\{i_1, i_4\},T_6)$
$=16+25+16 =57.$

Definition 4: A symptom is called a high utility element, if a particular symptom is occurring repeatedly, and is considered less useful, if less occurs based on a specific threshold, which is defined as minimum utility. If not, it is called a set of low utility elements.

Example 4:

Let there be a set of patients, identified with symptoms, with a minimum utility = 30

For the set of elements {i1, i4}, the utility is 57. It is greater than the minimum utility 30 specified by the user. Therefore, {i1, i4} is a set of highly useful elements.

Definition 5: The chances of a patient T_d to get the disease is denoted as TU (T_d) and defined as $u(T_d, T_d)$.

Example 5:

$TU(T_1) = u(i_1,T_1)+ u(i_3,T_1)+ u(i_4,T_1)$
$=(2X5) +(5X2)+(3X2)=26.$

Definition6: The weighted sum of symptoms against the patients in the dataset t X is the sum of the symptoms of all the related symptoms pertaining to X, which is specified as TWU(X) and defined as

$$TWU(X)= \sum_{X \subseteq T_d \wedge T_d \in D} TU(T_d)$$

Example 6:

$TWU(\{i_1, i_4\})=21+50=71.$

Definition7: An patient X is termed as highly prone to disease, if the weighted symptoms set , (HTWUI) is no less than minimum utility;

Example 7:

$TWU(\{i_1, i_4\})=21+50=71.$

The HTWUI value of item set $(\{i_1, i_4\})= 71$, which is greater than user specified minimum utility value 30.so $(\{i_1, i_4\})$ is a HTWUI.

Definition 8: Symptom Weighted Downward Closure Property (SWDC property): For any item set X , if X is not called as HTWUI , any super set of X is a low utility item set.

Example 8:

$TWU(\{i_1, i_4\})=21+50=71$

The HTWUI value of item set $(\{i_1, i_4\})= 71$, suppose that user specified minimum utility value =30 so $(\{i_1, i_4\})$ will not be a is HTWUI and all the super sets of $(\{i_1, i_4\})$ are high utility item sets.

A symptom set is an highly risky, if the diseased prone symptoms are greater than user specified minimum utility. Finding all HUI sets in the given database is called HUIM.

The most common and widely accepted method for finding the Parkinsons disease is Unified Parkinson's Disease Rating Scale (MDS-UPDRS) from Movement Disorder Society (MDS) .MDS-UPDRS is a questionnaire consisting of different parts concerning the progression of disease symptoms to be answered by the patients. It consists of question regarding the severity of the disease about symptoms like non-motor symptoms and cognitive symptoms(sleeplessness,headache,Memory loss etc.,), secondary motor related problems(Speech variation etc.,), Primary motor related Problems(like Muscle jerks,Walking weekness etc.,). Wether a person is effected with Parkinson's disease or not is represented by a column as status in the questionnaire.Each question from the MDS-UPDRS has 5 responses as 0 = normal (symptom not present), slight =1 , mild = 2, moderate = 3, severe = 4 . From the MDS-UPRDS questionnaire a symptoms data set, a database is created as symptoms represents attributes in the data table, and the responses becomes values of the corresponding attributes(as 0 to 4 based on the severity of the symptom).

3. DATASET

The MDS-UPRDS questionnaire data set is available in the Parkinson's Progression Markers Initiative (PPMI), a research organization that works to find solutions for degenerative diseases.. PPMI data MDS-UPRDS is a database composed of data from study sites for neurological disorders and movement disorder study in the United States, Europe, Australia and Canada, and the data set is available in CSV format. The data set consists of data from 405 patients, of whom 265 are men and 140 are women. The patients are aged between 33 and 84 years, with an average age of 61 years. There are a total of 1335 instances in the data set

4. SYMPTOMS

Various key observations with regard to the disease identification are notified and these indicators are considered as the biomarkers, these symptoms include:

- ✓ Tremor, also considered as shaking, typically begin in the segment of limbs, later on widen to hands and fingers.
- ✓ commemoration loss
- ✓ Muscle sensations
- ✓ Pains near temples
- ✓ restlessness
- ✓ deviation in speech
- ✓ alteration in writing skills

5. EXPERIMENTATION & RESULTS

A sample database containing the symptoms of P.D. is provided below. Each row, the table represents the symptoms of a sick person. The numerical value in the row represents the corresponding level or the severity of the symptom that can be considered as internal utility. The name of the symptom in the table is represented by a symptom name as well as a column number. The importance of the symptom is presented as a separate table considered as external utility.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Each symptom of a disease has a particular importance. For example, in many diseases, headache is a common symptom. Therefore, by considering the solitary symptom of headache alone, we cannot identify a particular disease. Some more symptoms will be there to identify that particular disease. The symptoms that are used to identify a disease should receive greater weight or importance compared to other symptoms.

Table 3: Symptoms of the Parkinson's Diseased Persons (internal utility)

Patient id Column no (0)	Speech Variation (S) (1)	Head Ache (H) (2)	Walking weakness (W) (3)	Muscle jerks (Mj) (4)	Memory loss (M) (5)	Sleeping Disorder (Sd) (6)	Status (st) (7)
P1	3	2	0	3	0	0	1
P2	2	0	0	4	2	0	1
P3	3	0	5	1	0	3	1
P4	1	0	3	0	1	2	0
P5	1	0	0	3	2	0	0
P6	1	2	0	4	0	0	1
P7	2	3	2	0	1	1	0
P8	0	0	0	0	0	2	0
P9	1	0	3	3	0	0	1
P10	5	0	0	4	0	0	0

Table 4: External Utility table

Name of the symptom	Speech Variation	Head Ache	Walking weakness	Muscle jerks	Memory loss	Sleeping disorder	status
Weight given to the symptom	15	3	1	12	5	3	10

The high utility item sets instituted with a least utility of 100 on this data set we have found the subsequent top HUI sets as follows

HUI sets produced with a minimum utility value =100

Table 5: High Utility Items sets on the above dataset

Sl. No	Item set	Utility value
1	6 3 1	118
2	6 1	108
3	3 1	100
4	2 7 4	116
5	2 7 4 1	176
6	2 4 1	156
7	2 1	111
8	5 4	104
9	5 4 1	149
10	5 1	120
11	7 4	208
12	7 4 1	252
13	7 1	175
14	4	252
15	4 1	396
16	1	270

The above table represents the high utility item sets generated from the sample dataset. In the item set column each number represents the column number of Table 3.From above example for item set(6,3,1) utility value is 118, which represents (Sleeping Disorder(6),Walking weakness(3),Speech Variation (1)) whose utility value is 116. Similarly Column 7 represents the people affected with the Parkinson's disease.

Table 6: Top High Utility Itemset Features

Top High utility Item sets	Utility Value
2 7 4	116
2 7 4 1	176
7 4	208
7 4 1	252
7 1	175

The item sets in the above Table 6 represents the symptoms associated with Parkinson's diseased people. Among them item set(7 4 1) has highest utility value of 208, which says that a P.D.effected person has significant symptoms of column 1 (Speech Variation) and column 4 (Muscle jerks).The significant features of a Parkinson's disease effected person are Speech Variation and Muscle Jerks. The same is showcased in the following bargraph-1

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

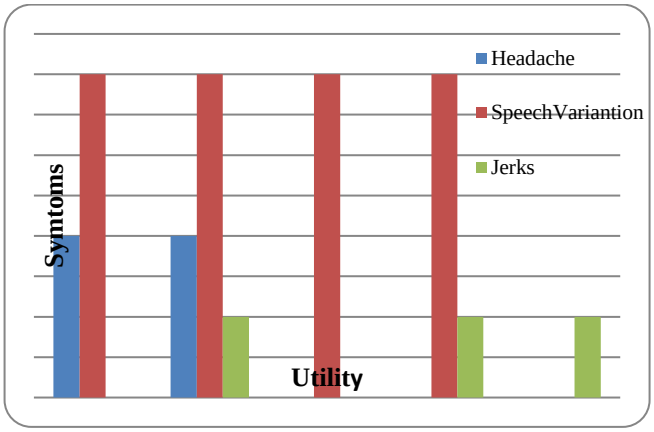

Fig:1 Graph showcasing the most significant symptoms

6. CONCLUSION

In this article a an ideology is projected to spot out the most noteworthy symptoms for identifying whether a particular symptom is to be interrelated with the Parkinson's disease or not and also this work helps to make out the most widespread symptoms that cause the P.D. This present article proposes that the key symptoms to be noted are jerks, allied with change in voice accent and head ache. The works presented help to recognize the syndrome at the near beginning stage such that the medical professionals take valuable steps in combating the disease.

REFERENCES

1. Bellou, V.; Belbasis, L.; Tzoulaki, I.; Evangelou, E.; Ioannidis, J.P. Environmental risk factors and Parkinson's disease: An umbrella review of meta-analyses. Parkinsonism Relat. Disord. 2016, 23, 1–9.
2. Berg, D.; Postuma, R.B.; Adler, C.H.; Bloem, B.R.; Chan, P.; Dubois, B.; Gasser, T.; Goetz, C.G.; Halliday, G.; Joseph, L.; et al. MDS research criteria for prodromal Parkinson's disease. Mov. Disord. 2015, 30, 1600–1611.
3. Babu GS, Suresh S. Parkinson's disease prediction using gene expression – A projection based learning meta-cognitive neural classifier approach. Expert Syst Appl. 2013; 40(5):1519–29. doi:10.1016/j.eswa.2012.08.070.
4. Rustempasic I, Can M. Diagnosis of Parkinson's Disease using Fuzzy C-Means Clustering and Pattern Recognition. SOUTHEAST Eur J SOFT Comput. 2013; 2(1):42–9. Available from: http://scjournal.com.ba/index.php/scjournal/article/viewFile/43/40
5. Abinaya S, Sivakumar R, Karnan M, Shankar DM, Karthikeyan M. Detection of Breast Cancer In Mammograms - A Survey. Int J Comput Appl Eng Technol. 2014; 3(2):172–8.
6. Shahbakhi M, Far DT, Tahami E. Speech Analysis for Diagnosis of Parkinson's Disease Using Genetic Algorithm and Support Vector Machine. J Biomed Sci Eng. 2014; 7(4):147– 56. doi:10.4236/jbise.2014.74019.
7. Defeng Wu, Kevin Warwick, Zi Ma, Jonathan G. Burgess, Song Pan, Tipu Z. Aziz " Prediction of Parkinson's disease tremor onset using radial basis function neural networks " Expert Systems with Applications 37 (2010) 2923– 2928.
8. Maria C Rodriguez-Oroz, Marjan Jahanshahi, Paul Krack, Irene Litvan, Raúl Macias, Erwan Bezard, José A Obeso "Initial clinical manifestations of Parkinson's disease: features and pathophysiological mechanisms" Lancet Neurol 2009; 8: 1128–39
9. Hartelius L. • Svensson P. "Speech and swallowing symptoms associated with Parkinson's disease and multiple sclerosis: a survey" FOLIA PHONIATR LOGOP. 1994;46(1):9-17.
10. Sofie Lundgren, Thomas Saeys, Fredrik Karlsson, Katarina Olofsson, Patric Blomstedt, Jan Linder, Erik Nordh, Hamayun Zafar, and Jan van Doorn "Deep Brain Stimulation of Caudal Zona Incerta and Subthalamic Nucleus in Patients with Parkinson's Disease: Effects on Voice Intensity " SAGEHindawi Access to Research Parkinson's Disease Volume 2011.
11. Hirsch, L.; Jette, N.; Frolkis, A.; Steeves, T.; Pringsheim, T. The incidence of Parkinson's disease: A systematic review and meta-analysis. Neuroepidemiology 2016, 46, 292–300.
12. B.Mouleswararao,Dr.Y.Srinivas, ,Towards Efficient Identification of Parkinson's Disease based on Frequent Pattern Mining and GMM, Jour of Adv Research in Dynamical & Control Systems, Vol. 10, No. 10, 2018.
13. Elbaz, A.; Carcaillon, L.; Kab, S.; Moisan, F. Epidemiology of Parkinson's disease. Rev. Neurol. 2016, 172, 14–26, doi:10.1016/j.neurol.2015.09.012.
14. Drotar, P.; Mekyska, J.; Rektorova, I.; Masarova, L.; Smekal, Z.; Faundez-Zanuy, M. Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinson's disease. Artif. Intell. Med. 2016, 67, 39–46.
15. Brabenec, L.; Mekyska, J.; Galaz, Z.; Rektorova, I. Speech disorders in Parkinson's disease: Early diagnostics and effects of medication and brain stimulation. J. Neural Transm. 2017, 124, 303–334.
16. De Stefano, C.; Fontanella, F.; Impedovo, D.; Pirlo, G.; di Freca, A.S. Handwriting analysis to support neurodegenerative diseases diagnosis: A review. Pattern Recognit. Lett. 2018, in press.
17. Thomas, M.; Lenka, A.; Kumar Pal, P. Handwriting Analysis in Parkinson's Disease: Current Status and Future Directions. Mov. Disord. Clin. Pract. 2017, 4, 806–818.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)