A Data Mining Tool for Prediction of Suicides among Students

Omprakash L. Mandge

Mumbai Educational Trust
MET Complex, Bandra Reclamation
Mumbai
India
olm.deep@gmail.com

Abstract

The inability to handle pressure and meet parent's expectations is giving rise to suicides among students in high schools and colleges. This is leading to the loss of talent which can be useful in nation building. There is increasing need for providing valid means of determining which students are at risk for suicide. The ability to predict suicidal behaviour is still relatively poor, but identification of personality traits and behaviours in individuals can be helpful to minimize the number of students committing suicides. In this paper we study on implementation of a counselling system to predict suicidal tendencies and depression among the students. We analyze the different warning signs (observable and non observable) such as interpersonal communications, interpersonal relations, classroom behaviour etc. and apply the data mining algorithms to generate the results. We collect data about different students and then design a gradation system on the basis of collected results. Using this gradation system we determine the students at a high risk.

1. Introduction

Many young lives are lost everyday as a result of self inflicted deaths, suicides. The increase in suicidal behaviour among youth has become a major concern in the recent past. It is albeit painful to accept this trend, but it is triggering a major concern in India. Each youth death by suicide causes sufferings to families, friends, communities and practitioners. Ultimately, it leads to loss for the nation, as the future of India is the youth.

Suicide has been described as the end of a continuum that begins with suicidal ideation, continues with planning and preparing for suicide, and ends with threatening, attempting, and completing suicide (Kuchar, Potter, Powell, & Rosenberg, 1995). Although some young people make impulsive attempts, many more have suicidal thoughts and engage in behaviours along this continuum before attempting suicide or without ever attempting suicide.

It is estimated that over 100,000 people die by suicide in India every year. India alone contributes to more than 10% of suicides in the world. The suicide rate in India has been increasing steadily and has reached 11.2 (per 100,000 of population) in 1999 registering a 78% increase over the value of 1980. In year 1999, 5% of all the suicides were committed by students. In 2006, 5,857 students or 16 a day committed suicide across India due to exam stress. And these are just the official figures. This clearly suggests that there is tremendous increase in suicides among youth in India and there is a need to prevent as well as control suicides.

As a serious public health problem, it demands our attention, but its prevention and control, unfortunately, is a very difficult task. Researchers and psychologists claim that a suicide is preventable, but involves a lot of activities ranging from the provision of the best possible conditions for bringing up our children and youth, through the effective treatment of mental disorders, to the environmental control of risk factors. Appropriate dissemination of information and awareness-raising are essential elements in the success of suicide prevention program.

Suicide is preventable and controlled by observing the students over a period of time. Youth who are contemplating suicide frequently give warning signs of their distress. Parents, teachers, and friends are in a key position to pick up on these signs and get help. The major concern is how to observe these warning signs and predict the students at high risk. For accurate prediction, we need collect student data on periodic basis and set the parameters to measure the signs. Data mining can be very helpful to achieve this.

ISBN 978-93-82338-79-6

Data mining is the fundamental stage inside the process of useful and comprehensive knowledge, previously unknown from large quantities of stored data in different formats with the objective of improving the decisions of companies, organizations or institutions where the data have been gathered. The overall process of data extraction and information generation is known as Knowledge Data Discovery KDD [3, 4].

In this paper we apply the process of data mining in order to analyze the student data and on the basis of analysis we provide means to predict suicidal behaviours. We apply different learning methods and identify the most accurate methods. Then these methods can be implemented in the proposed tool.

This paper is organized as follows: In section 2 we explain the data mining virtuous life cycle and propose the structure of tool. Section 3 discusses the business objectives for the system and conversion of those objectives into data mining objectives. In subsequent sections we concentrate on data collection, transformation and model creation. The final section emphasizes the model evaluation and selection.

2. The Virtuous cycle of Data mining

The complexity of the data mining process is well captured by CRISP-DM methodology [5]. On the basis of this methodology we have formalized the data mining virtuous life cycle [6]. Mainly the life cycle consists of 1) Translate business opportunity (problem) into DM opportunity (problem) 2) Select appropriate data 3) Get to know the data 4) Create a model set 5) Fix problems with the data 6) Transform data to bring information to the surface 7)Build models 8) Assess models 9) Deploy models 10) Assess results 11) Begin again

The first step in any data mining process is to set up the business objectives. These objectives include: Understanding the student behaviours, identifying the students at high risk, Analyze the previous year's student data, help to predict the future student performance etc. Objectives like these are of major interest for the identification in any institution. So these objectives can work as a set of generic objectives for the initial analysis.

3. Data Collection

Once the objectives are defined, we need to identify and collect the data to be used. In our case we need two types of data: Internal (collected by the student management systems) and external data (such as demographic details of the students, area, culture, attitude of the people towards education etc). The internal data can be collected very easily, but it's very difficult to get the external data.

Besides this, we need to identify the behavioural changes and factors responsible for suicidal ideation in the student. Stillion, Mcdowell and May (1998), proposed a model as shown in fig 1 known as "Suicide Trajectory model". This model suggests that there are four major categories of risk factors that contribute suicidal behaviour at every age: biological, psychological, cognitive, and environmental. As the arrow indicates, each of these categories of risk factors may directly influence suicidal ideation and may affect other categories of risk factors.

One of the major issues here is how to identify the risk factors and how to collect them. Based on the literature review and available resources we think the following factors can be very helpful in prediction of suicidal behaviours.

3.1 Classroom Behaviour

- Marked decline in school performance and levels achieved
- Skipping classes and opting out of school activities generally
- Poor concentration, sleepiness, inattentiveness
- Unusually disruptive or rebellious behaviour
- Death or suicide themes dominate written, artistic or creative work
- Loss of interest in previously pleasurable activities
- Inability to tolerate praise or reward.

3.2 Interpersonal Behaviour

- Giving away prized possessions
- Sudden change in relationships, e.g. exhibiting disruptive behaviour.
- Withdrawing from friends and social involvement.
- Not wanting to be touched by others.

3.3 Other Behavioural Signs

- Apathy about dress and appearance.
- Sudden change in weight
- Running away from home.
- Risk-taking and careless behaviour
- Accident proneness.
- Sudden and striking personality changes and changes in mood.
- Overt signs of mental illness e.g. Hallucinations.
- Loss of sense of humour or sudden compulsive joking.
- Sleeping pattern changes.
- Self-mutilation behaviours.
- Noticeable increase in compulsive behaviour.
- Development of extreme dependency
- Sudden happiness after a prolonged period of depression.
- Impulsive tendencies
- Depressive tendencies
- Unrealistic expectations held of self.

ISBN 978-93-82338-79-6

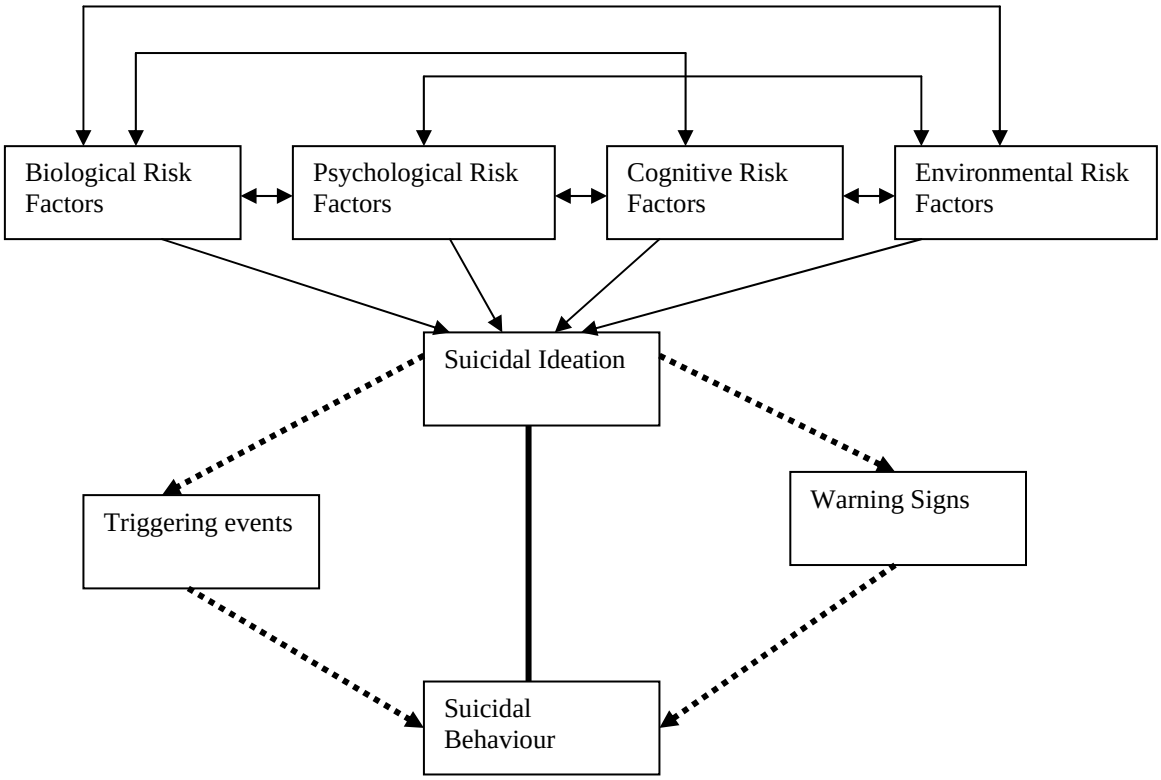

Fig 1. Suicide Trajectory Model (Stillion, Mcdowell and may 1998)

3.4 Interpersonal Relations

- Involvement in physical violence
- Inability to relate well to peers.
- Sexual promiscuity.
- Inability to enjoy or appreciate friendships or to express affection openly.
- Mood swings and occasional outbursts.
- Feelings of worthlessness, being a burden or having let parents or others down.
- Feelings of guilt, failure, having no control over their lives.

As per the above mentioned factors, the data related to classroom behaviour can be collected very easily as almost all schools and colleges are using sophisticated information systems. The data can be readily available and needs very less processing. But, a major issue here is how to record and collect the data related to Interpersonal behaviour, other behavioural signs and Interpersonal relations. It is very difficult to observe and identify these traits in students. The identification of such traits is a complex activity and requires a lot of attention as well as continuous monitoring of the student.

To collect the data in relation these traits we have prepared a Questionnaire by consulting different psychologists and doctors. We have prepared a set of 50 questions most of which are multiple choices. We assign weights ranging from 1 to 5 for each answer and then collect and store these weights in proposed tool. These values are stored along with student demographic data and classroom performance data. The entire data forms the basis for analysis.

4. Data Pre-processing

The next step in the data mining method is the pre-processing of the data collected in the model set. The pre-processing comprises the stages of identifying, collecting, filtering and aggregating data into a format required for the data warehouse. Most of these traditional data pre-processing tasks are not necessary for the data collected from school or college administration systems, as the data is available in sophisticated software systems.

The Data collected from the questionnaire and interviews with parents as well as students, needs a lot of pre-processing activities. This day may contain missing values and may not be formatted properly. We apply the different data mining pre-processing activities such as reconciliation, aggregation, regression, outline analysis on the collected data and store it in a single file format. After the pre-processing activities, the data can be used for the further analysis. Then we apply the different learning methods on stored data.

ISBN 978-93-82338-79-6

5. Learning Methods

Once the data have been properly filtered, cleaned and transformed, we can proceed with the induction of the prediction models. For this purpose, we employ the suite WEKA [10], and we make our modules work with it. This suite integrates many of the most known learning techniques, as well as, several pre-processing and post-processing tools. Additionally, WEKA has been released as open source, so, if it is required, we can adapt this software for our particular requirement

The key point for using WEKA is the proper construction of the minable view in such a way that could be directly used by the learning methods. A standard format (arff) has been defined as a data and model file repository in WEKA. So, the idea is to generate the data in this format, and in this way we can employ all the different leaning techniques integrated in this suite. In our case, we generated different minable views for some areas of our interest on the basis of sample data.

The data for sample study is collected from the prepared questionnaire and the weights for each question are calculated and collected automatically in the tool. For the study, we have collected data from the 315 students' including 195 males and 120 females. The average age of these students ranged between 13 to 21 years, with a mean age of 17 years (Standard Deviation of 2.03). With this initial data we have prepared different minable views, although we show only one as bellows

Attribute	**Description**
Student ID	Students ID
thisPerc	Result of the student this year
LastPerc	Result of the Student Last year
His	Student History
Age	Age of the Student
Attendance	Number of classes the student attends
Behaviour	Weight for the current behaviour
Total	Addition of all the weights

Table 1. A Minable View

With this initial minable view, we used different learning methods included in WEKA: LinearRegression, LeastMedSq, SMOreg, MultilayerPercepton, Kstart, LWL, Tree DecissionStump, Tree M5P and IBK. We used 10-fold cross validation in the experiments. The method which obtained the best results was the linear regression and tree M5P.

Besides this we have designed a Gradation System and apply the Baysian Classifier to produce the results. The Results clearly shows the students at suicide risk.

6. Conclusions

Prediction of suicidal behaviour is still relatively poor. In this paper, we have analyzed the adequacy of designing specialized modules for data mining for Suicidal behavior prediction among students and we have also identified which are the stages in the KDD process which could be used and implemented for the design of tool. The success of this project could turn data mining into an available technology to identify the students at high risk of suicide.

On the other hand, the accuracy of the prediction depends on the collected data. The data needs to be updated on periodic basis.

The tool will definitely help to identify the students at risk of suicide.

8) References

1 http://www.maithrikochi.org
2 Kachur S, Potter L, James S, Powell K. Suicide in the United States, 1980-1992. Violence Surveillance Summary Series, no. 1. Atlanta: CDC, National Center for Injury Prevention and Control, 1995.
3 SASInstitute: http://www.sas.com/feature/4qdm/whatisdm.html
4 J. Luan. Data mining and knowledge management in higher education: Potential applications. In Annual Forum for the Association for Institutional Research, Toronto, Ontario, Canada, June 2002
5 CRISP-DM, www.crispdm.org
6 Mastering Data Mining, Ron Norman
7 Assessing suicide risk in children: guidelines for developmentally appropriate interviewing. Publication Title: Journal of Mental Health counseling Barrio, Casey A
8 World Health Organization. Prevention of suicide: guidelines for the formulation and implementation of national strategies. Geneva: World Health Organization, 1996.
9 National Crime Records Bureau (NCRB) report for the year 1999 and 2006
10 www.weka.org

ISBN 978-93-82338-79-6