A Prediction Model for Criminal Levels Specialized in Brazilian Cities

Marcelo Damasceno de Melo[1,2], Jerffeson Teixeira[2], and Gustavo Campos[2]

[1] IFRN - Campus Macau, R. das Margaridas, 350, Macau-RN, Brazil
http://www.ifrn.edu.br
[2] Universidade Estadual do Ceará, Av. Paranjana, 1700, Fortaleza-CE, Brazil
{marcelodamasceno,jeff,gustavo}@larces.uece.br
http://www.uece.br

Abstract. The increase in violence around the world is becoming a major problem, causing severe damages to society: material, social and physical ones. The Government needs effective tools to fight against crime, and therefore, some tools are necessary to assist in the prevention of further crimes, in the allocation of its resources and visualization of geographic areas with high crime concentrations.

This paper proposes a model of data mining, predicting criminal levels in urban geographic areas. The model was proposed to work using Brazilian data, specifically criminal and socio-economic ones. This work shows the approach proposed to face the problems of this social phenomenon, as a unified process to build a system which can able to help decision managers to fight and prevent crime.

To validate the proposed procedure it was used as a case study. Using the crime and socioeconomic data of the Metropolitan Region of Fortaleza - Brazil (RMF). The case study proved that the process is useful and effective in building a predictor of criminal levels. The model achieves 70% of accuracy using an innovative method and heterogeneous data sets.

Keywords: Prediction Model, Criminal Levels, Data Mining Model, Brazilian Crime, Predicting Crime.

1 Introduction

The increase in violence in recent years has been the object of study of many researchers. Governments and society in general have problems with the inconvenience caused by this phenomenon (violence). Each year, governments spent millions of dollars combating the violence, providing equipment, training and purchasing tools to assist the police work.

Each crime can raise a lot of data, such as: date, time and place of the crime, *modus operandi*, crime type and socio-economic status of the victim. This type of data may facilitate the use of data mining tools such as prediction. Such data are important because they provide a computer system able to predict

H. Jahankhani et al. (Eds.): ICGS3/e-Democracy 2011, LNICST 99, pp. 131–138, 2012.

the occurrence of crimes and even listing all the variables that affect the event (crime).

The crime prediction is often used for predicting future places where crime will occur [1,2,3]. Several theories develop activities within the study of criminal behavior, like: routine activities, *hot spot* [4] and crime ecology.

Data mining is one of the steps in a process of discovering patterns embedded in the data [5]. Data mining has been important because it is a powerful tool to extract valuable information found in a database. There are applications in some areas such as fraud detection, lifting profile, marketing and monitoring.

Our work is a proposal of a prediction model using criminal and socio-economic data. The model deals with all steps to build a system to predict criminal levels using criminal and socio-economic data. Criminal data are information about date, local and informations about the crime and socio-economic data are information about economic and social variables in a city, like salary or number of schools. We define the criminal levels as a measure of danger in a certain demographic region. One of the most interesting aspects of our work is the use of socio-economic aspects, because these factors have great relevance in the crime occurence and were not taken into account in several studies.

2 Prediction

Every corporation plan their day-to-day actions based on data generated in their actions. Data collected must illustrate their experiences and demonstrate the rights and wrongs committed daily. Quantitative forecasting models basically use historical data to detect patterns and estimate them in the future. Thus, the acquisition of tools of this kind should be seen as an organizational gap, add support for the decisions made by managers.

The act of predicting can be defined as obtaining an accurate answer on a question that should happen in the future, based on the past. The future shall be understood as a scenario or a situation never experienced by a corporation or something you want to pursue. Thus, the predictions must be conducted in completely independent variables based on data from past and present stored in databases and experience of managers and other professionals involved.

A process that determines the steps to build a prediction systems must be followed because it facilitates the development of the adjustments and the resolution of problems arising in the implementation of a forecasting technique [6]. One of the purposes of this work is to develop a model based on a process for developing a prediction application to criminal levels, using machine learning algorithms, which can be applied to any Brazilian city.

3 Prediction Model

The process of Knowledge Discovery in Databases (KDD) is not a simple task [7]. The information obtained from the application of this process is not just coming from the direct application of learning algorithms, but also the understanding

of the business, collecting, cleaning and processing of data, post-processing and visualization of knowledge.

The approach proposed in this paper serves as a guide to conduct crime level prediction in a Brazilian city, with little adjustment for application in any cities. The prediction process must use social, economic and criminal data, because crime is a phenomenon that can be explained using these characteristics. The definition of these data was based on characteristics of the crime phenomenon observed in the literature. The process will address traditional and new tasks such as task-specific prediction of crimes using Brazilian data.

The definition of a specialized model for the problem of crime prediction makes the resolution of the problem easier to achieve using data mining. The model will deal clearly and objectively with all the steps in order to provide convenience and security in the results.

The Figure 1 shows the steps that must be executed. Each step must be executed in order that it appears. There are single and double connections between steps. Single connections mean there is just a way of execution, without return; and double connections mean there is a way to return to the previous step. After the Evaluation task, there are two connections, one to the Distribution step and another one to Bussiness Understanding. It means if the evaluation is bad, there is a way to restart all the process. Thus, the process allows a way to fix some problems that were not view previously. If the evaluation is good, the process guide to the Divulgation task. All the steps shown in the Figure 1 were implemented in the predictive model.

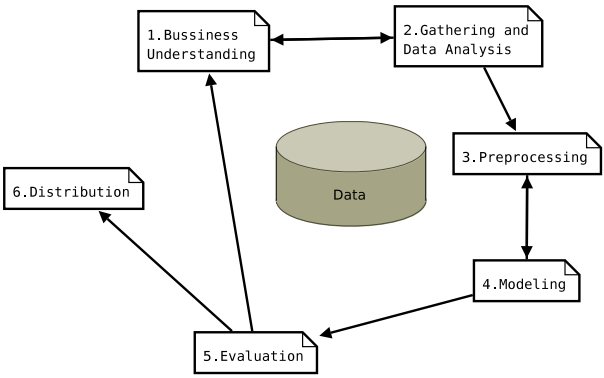

Fig. 1. Steps of the prediction model

Each model step below is described from the characteristics experienced in the case study using the proposed model. The case study used socio-economic and crime data of the Metropolitan Region of Fortaleza (`MRF`). `MRF` is a big region with 13 cities and more than 2 milions citizens.

3.1 Bussiness Understanding

This step performs an understanding of crime in the city that will be used in the study. It must consider criminal, social, demographic and economic characteristics. It is essential to raise small details like date, time, location and type of crime committed. Besides raising data on crimes committed, it is important to collect social data for the region such as population density, illiteracy, rates of life quality, sport and culture, numbers of people distinguished by sex and age.

In addition to understanding the criminal, social and economic issues, it is necessary to define the technical objectives to be achieved. As the proposed model is focused on predicting criminal levels, it was decided to use five criminal levels calculated from the number of crimes committed grouped by criminal and social/economic variables. The levels are very low, low, medium, high and very high. Thus, the predictive method will consist of five models that will say what level a given region belongs. Each model will be specialized in predicting a certain criminal level.

Besides defining the data that must be used and the data mining goal that the analyst must achieve, this step has a task that guides what minimum accuracy must be get to the project has to be accept. We suggest that the true rate must be defined by a crime specialist at the demographic region that is being studied. This work was not supported by a crime specialist, so we accept a true rate superior than 50%.

3.2 Gathering and Data Analysis

This step will deal with how to obtain and how to analyse the data. As it was said before, we suggest that the model uses crime and socio-economic data. Below is described how we obtained the data used in our case study.

The data crime used in the case study were obtained through the Department of Public Safety and Social Defense of the State of Ceará(`SSPDS-CE`). The information refers to crimes committed between 2007 and 2008. The available data contain information about time, location, type and subtype of the crime.

The socio-economic data obtained are related to the Census occurred in 2000 conducted by the Brazilian Institute of Geography and Statistics (`IBGE`). The data contain all the variables studied in the Census grouped by geographic areas known as Data Expansion Areas (`DEA`). Several districts may belong to the same `DEA`, because the definition of so large a `DEA` is dependent on population variables. So the extension of the `DEA` depends on the number of citizens for example.

3.3 Preprocessing

It was necessary to integrate criminal and socio-economic databases to become only a table that contained all the collected data. To integrate the databases was used the attribute `DEA`, originally belonging to socio-economic database. The transformation performed in the crime data creates a `DEA` attribute too. After

the integration is completed, the dataset has 69 socio-economic and criminal attributes.

As it was decided in the bussiness understading step, our prediction model will be composed of 5 models, each one will predict a specific criminal level. Thus, it must divide the database (flat table) into five subdivisions, setting each example to their criminal level.

The regions with the lowest amount of crime received the very low crime level using all the criminal, social and economic variables.

The very high level was determined for the regions with the highest crime numbers using all the variables previously cited. The intermediate levels (low, medium and high levels) were calculated using a scale. The minimum (min) and maximum (max) value of this scale is lowest and the highest crime numbers respectively. Each step (s) at a scale was determined by ($max - min/5$).

Thus, 5 different sets are generated, each one representing each level. Each dataset has the same number of instances and attributes of the original set.

To improve the accuracy and running time it is needed an attribute selection. We suggest the use of the Ranker search method using the Information Gain as a measure of evaluation [5]. The attribute selection was applied and it selected the 15 best attributes of each dataset, according to the evaluation measure used. Different attributes were extracted for each dataset, thus, it was necessary to standardize the attributes that would be used. It was chosen the seven best attributes of each execution of the selection algorithm. Now each dataset only has 35 attributes, seven attributes selected from each dataset.

The set of attributes selected by the selection algorithm does not contain attributes so important as the city, day, month and year of crime occurrence. As the temporal attributes are important for understanding the criminal dynamics, it was decided to build a new dataset containing the attributes selected by the selection algorithm plus the temporal attributes excluded. So we have two predictive models, one constructed with the selected attributes and another one with the selected attributes plus the excluded ones.

3.4 Modeling

Several machine learning algorithms were tested, but the Neural Networks were chosen [5] due to the results achieved and the constraints imposed by the data. For each data set, a network was trained, where the network will be used to predict the crime level of a certain region in a near future. We have two predictive models, composed of five Neural Networks each. Remember, the first model was trained using the first set of attributes and the second on using the same attributes of the first set plus the temporal attributes excluded by the selection algorithm.

3.5 Evaluation

With the use of trained neural networks it was obtained an accuracy (hit rate) over 70% and mean error and mean square error of at most 0.36 and 0.47,

respectively. As it was defined at the bussiness understanding step, the accuracy must be at least 50%, so the 70% of accuracy was a good result.

3.6 Distribution

It was utilized plots to show the predictions. After a question is put on the system, the system outputs a plot showing its answer. The plot has two lines to represent the prediction and the model confidence. The model confidence was calculated using the neuron activation values of the exit layer of neural net. The black line shows the predicion made by each model. When the value is 1, it means that region has the level represented in the abscissa, and 0 otherwise. The grey line determines how much of confidence each model has in its prediction.

Thus, the plot displayed in Figure 2 shows that the predictors that represents high and very high levels say that instance belongs to their levels with 65% and 60% of confidence respectively. All other models deny the participation of the instance in their levels.

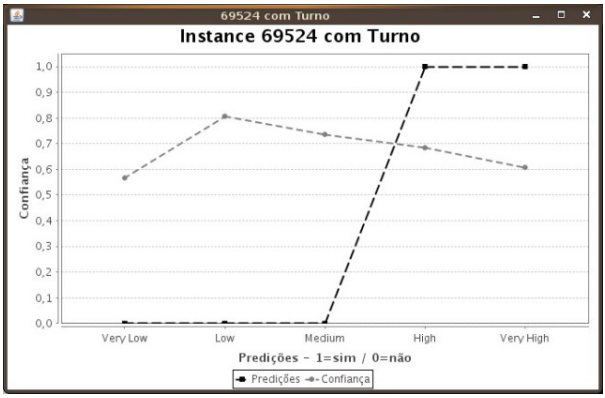

Fig. 2. Prediction of a negative instance of medium level belongs to the second data set

The chart in Figure 3 shows that all models predict the instance as negative. The question was made for the first model, the model trained with the first data group, the group with 36 attributes. A greater confidence for the prediction was made by the model that predicts the high level and the smallest on for the model that predicts the very low level. The prediction was not so wrong, it was wrong for the class, but its confidence in asserting such prediction for the very low level was low. Thus, we can conclude that region may have very low criminal level. We will use the second model to ensure the prediction made by the first model.

The figure 4 displays the prediction using the same question used in the previous example, but now using the attributes of the second group of data. Like the previous result, all models predict the instance as negative for the criminal

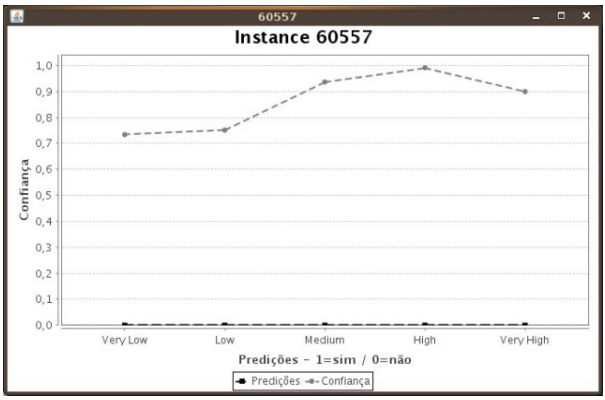

Fig. 3. Prediction of a first data set instance with positive very low level

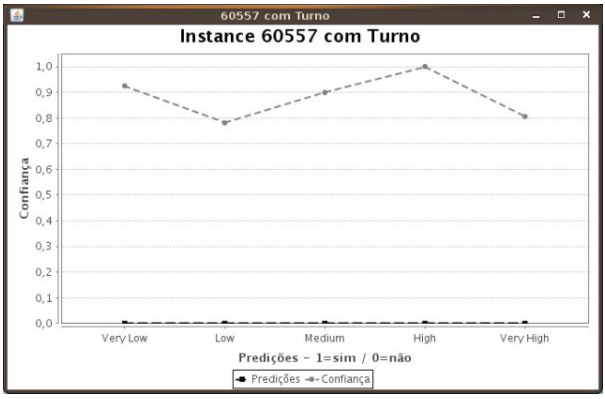

Fig. 4. Prediction of a second data set instance with positive very low level

level that it represents. The lower confidence in the statement was made by the model that predicts the low level and the highest value was for the model that predicts the instances with high level. Unfortunately the second model showed a high confidence in denying the very low level, however denying the low level with low confidence. From the results of the two predictors we can conclude that the criminal level of the region is on the threshold between low and very low, based on the result of the first and second models. Thus, it can predict a trend in the growing of crime, from the very low level to low level. Besides predicting the criminal level, our model can show trends of growth or decline in violence.

Like the case before, there are cases in which the predictors have undecided results or low confidence in their results, for this, it is used the other model to support the results of the first executed model. Thus, a system uses two different predictive models, in which the analyst can use them to support their decisions.

4 Conclusion and Further Works

The model defined is made as a tool for help decisions makers using predictions. Its predictions were based on criminal levels using data from Brazilian geographic areas. Besides being possible to implement it in any city in Brazil, because the necessary data can be collected locally through its security secretary and IBGE. Our model may be extended for other cities for differents countries just doing few adjustments. We believed that the socioeconomic characteristics of a region influence local crime, this hypothesis is proven in our case study. Any industry can use our intelligence model to build an agile and reliable system to predict criminal levels.

For the case study it was used the Fortaleza Metropolitan Region. We got an accuracy better than 70% and a distinct method of analyzing the result. Using two models to reinforce the results obtained by the predictors.

As a future work, we will apply the built model using the latest data. Besides the evaluation, we suggest the use of a map system GIS for easier viewing of results. New learning algorithms can be used and may provide superior results to those found using Neural Networks.

References

1. Mitchell, M., Brown, D., Conklin, J.: A Crime Forecasting Tool for the Web-Based Crime Analysis Toolkit. In: IEEE Systems and Information Engineering Design Symposium, SIEDS 2007, pp. 1–5 (2007)
2. Henderson, M., Wolfers, J., Zitzewitz, E.: Predicting Crime. Arizona Law Review 52, 15–173 (2010)
3. Mohler, G., Short, M., Brantingham, P., Schoenberg, F., Tita, G.: Self-exciting Point Process Modeling of Crime. Journal of the American Statistical Association 106(493), 100–108 (2011)
4. Furtado, V., Ayres, L., de Oliveira, M., Vasconcelos, E., Caminha, C., D'Orleans, J., Belchior, M.: Collective Intelligence in Law Enforcement - The WikiCrimes System. Information Sciences 180, 4–17 (2009)
5. Witten, I.H., Frank, E.: Data Mining: Pratical Machine Learning Tools and Techniques, 2nd edn., Diane Cerra. The Morgan Kaufmann Series in Data Management Systems (2005)
6. Mahmoud, E., DeRoeck, R., Brown, R., Rice, G.: Bridging the Gap between Theory and Practice in Forecasting. International Journal of Forecasting 8, 251–267 (1992)
7. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communications of the ACM 39, 27–34 (1996)