修士論文

Automatic Detection of Repressed Anger from Text Messages

20〇〇年〇月〇〇日

(Replace with publication date)

学籍番号　16906291

ヘスス･マリア･セスマ･ソランセ

指導教員　湯川　高志

長岡技術科学大学大学院工学研究科

経営情報システム工学専攻

平成　　年　　月　　日

専攻名	経営情報システム工学	学籍番号	16906291
申 請 者 氏 名	ヘスス･ マリア･ セスマ･ ソランセ		
指導教員氏名	湯川高志　教授		

審査委員主査　○○　○○　教授

審査委員　○○　○○　准教授

審査委員　○○　○○　准教授

審査委員

審査委員

専攻主任印

論　文　要　旨

論文題目	Automatic Detection of Repressed Anger from Text Messages

The growth in the usage of social media, microblog and review platforms has resulted in significant increase of the access to short text messages that reflect individuals' opinion and feelings that enable Natural Language Processing. Detecting people's emotions has a wide range of applications such as producing systems that automatically measure the satisfaction of customers and thus can help companies to improve their products or services. This research project focuses on detecting anger, an emotion that is relative difficult to detect compared to other sentiments due to the usage of linguistics figurative language techniques, such as irony, that intends to communicate the opposite of what it is literally said.

To this purpose, a review of the state of the art has been made and an experiment using machine learning techniques has been conducted in an open social network like Twitter.

修士論文

Automatic Detection of Repressed Anger from Text Messages

20〇〇年〇月〇〇日

(Replace with publication date)

学籍番号　16906291

ヘスス･マリア･セスマ･ソランセ

指導教員　湯川　高志

長岡技術科学大学大学院工学研究科

経営情報システム工学専攻

Contents

List of Figures

List of Tables

Chapter 1

Classification Techniques

The aim of this chapter has two purposes. The first one, is to make an introduction of the basic concepts of classification, which is essential for the detection of repressed anger. The second one, is to explain how the algorithms used in this study work.

1.1 Fundamentals of Classification

According to [11], classification can be defined as the task of predicting an outcome from a given input. This outcome is produced by the process of mapping a group of characteristics present in the input to a certain category. In other words, it consists in assigning objects (the input) to one of several predefined classes (the outcome) [10]. Examples of classification can be found in everyday life, such as e-mail spam detection, news classifiers, Optical Character Recognition (OCR), animal kingdom classification (see Figure 1.1), among many others.

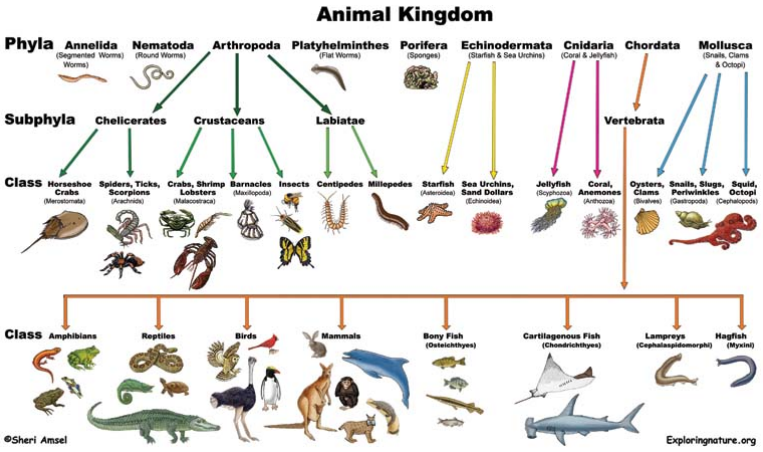

Figure 1.1: Classification of animals. The image is extracted from Exploring Nature.

The input data for a classification task is composed by a collection of records, the dataset. In the same time, each record, also known as an instance, is composed by a set attributes. From all these attributes there is one considered special, which

is called the target attribute or the class label. Regular attributes can be both discrete or continuous values. For the values signed for the class label, however, they must be discrete. This characteristic is what distinguishes classification form regression. Table 1.1 shows a sample dataset for animal classification into the following categories: amphibian, bird, fish or mammal.

Common Name	Hair	Feathers	Eggs	Milk	Aquatic	Legs	Class Label
antelope	Yes	No	No	No	No	4	mammal
catfish	No	No	Yes	No	Yes	0	fish
dolphin	No	No	No	Yes	Yes	0	mammal
dove	No	Yes	Yes	No	No	2	bird
duck	No	Yes	Yes	No	Yes	2	bird
elephant	Yes	Yes	No	Yes	No	4	mammal
flamingo	Yes	Yes	Yes	No	No	2	bird
frog	No	No	Yes	No	Yes	4	amphibian
fruit bat	Yes	No	No	Yes	No	2	mammal
gull	No	Yes	Yes	No	Yes	2	bird
herring	No	No	Yes	No	Yes	0	fish
kiwi	No	No	Yes	No	No	2	bird
lark	No	Yes	Yes	No	No	2	bird
lynx	Yes	No	No	Yes	No	4	mammal
mole	Yes	No	No	Yes	No	4	mammal
mongoose	Yes	No	No	Yes	No	4	mammal
newt	No	No	Yes	No	Yes	4	amphibian

Table 1.1: Animal kingdom dataset.

Tan Pang-Ning et al. propose a more mathematical definition of classification stating that it is the process of learning a target function f, also known as classification model, that maps each attribute set x to one of the predefined class labels y.

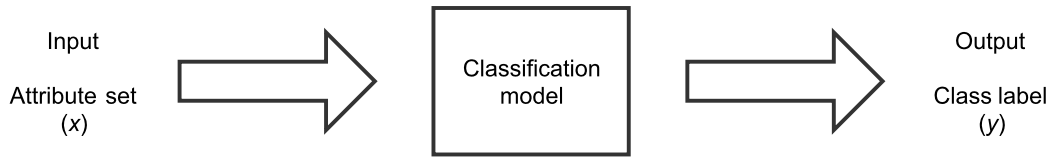

Figure 1.2: Classification as a task of mapping a set attributes x into its fitting class label y.

A classification model is useful for the following purposes [10]:

- **Descriptive Modeling:** Since a classification model presents the main features of the data, it can serve as an explanatory tool to distinguish between instances of different categories [9].

- **Predictive Modeling:** A classification model can also be used to predict the class label of an unknown new instance. As shown in Figure 1.2, a classification model can be represented as a black box that automatically assigns a class label to an instance by providing its attribute set.

It is important to remark that classification techniques perform their best when used for predicting or describing datasets which its class label is binary or nominal, Since they no consider properties such ordinality or the implicit order among the categories, they become ineffective with ordinal class labels [4].

1.2 General classification problem solving

For general classification problems solving, popular techniques consists on a process that starts with building classification models from a sample dataset [12]. Each technique depends on a learning algorithm witch is in charge of generating the classification model. A good model should define the relationship between the input attribute set and its belonging category that suits the best. Therefore the model should be valid for both, the sample data used to generate the model and also for new unknown instances. Among popular classification techniques Support Vector Machine (SVM), Neural Networks (NNs), Naive Bayes or Decision Trees (DTs) can be found [5].

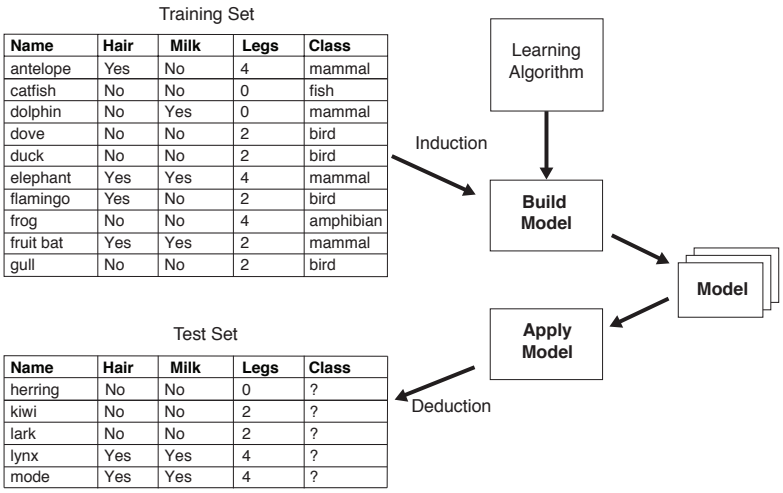

Figure 1.3: General approach for classification model building and new instance category prediction.

As shown in the Figure 1.3, to solve a classification problem a sample dataset must be provided as a training set. This sample is used to build the classification model according to the learning algorithm. After the model is built, it is applied to unlabeled dataset, also called the test set, to predict the categories of each instances of the records. To measure how good the model is, there is only need to count the number of instances have been correctly and incorrectly classified from the test set. Usually, to represent system's performance values, a confusion matrix is used [6].

		Predicted Class	
		$Class = Yes$	$Class = No$
Actual Class	$Class = Yes$	a	b
	$Class = No$	c	d

Table 1.2: Confusion matrix of a binary classification.

As a example, Table 1.2 represents the confusion matrix of a binary classification problem.

1.3 K-Nearest Neighbor

1.4 Neural Networks

References

[1]Laura Auria and Rouslan A Moro. "Support vector machines (SVM) as a technique for solvency analysis". In: (2008).

[2]Robert Berwick. "An Idiot's guide to Support vector machines (SVMs)". In: *Retrieved on October* 21 (2003), p. 2011.

[3]Gavin C Cawley and Nicola LC Talbot. "On over-fitting in model selection and subsequent selection bias in performance evaluation". In: *Journal of Machine Learning Research* 11.Jul (2010), pp. 2079–2107.

[4]Eibe Frank and Mark Hall. "A simple approach to ordinal classification". In: *European Conference on Machine Learning*. Springer. 2001, pp. 145–156.

[5]GV Garje et al. "SENTIMENT ANALYSIS: CLASSIFICATION AND SEARCHING TECHNIQUES". In: (2016).

[6]Howard Hamilton. *Confusion Matrix*. 2000. URL: `http://www2.cs.uregina.ca/~dbd/cs831/notes/confusion_matrix/confusion_matrix.html` (visited on 10/21/2016).

[7]Chih-Wei Hsu and Chih-Jen Lin. "A comparison of methods for multiclass support vector machines". In: *IEEE transactions on Neural Networks* 13.2 (2002), pp. 415–425.

[8]Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. "A practical guide to support vector classification". In: (2003).

[9]David Madigan. *Descriptive Modeling*. 2002.

[10]Tan Pang-Ning, Michael Steinbach, Vipin Kumar, et al. "Introduction to data mining". In: *Library of congress*. Vol. 74. 2006.

[11]Fabricio Voznika and Leonardo Viana. *Data Mining Classification*. 2007.

[12]Ian H Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

Acronyms

DT Decision Tree. 3

NN Neural Network. 3

OCR Optical Character Recognition. 1

SVM Support Vector Machine. 3