Auto Classifier

Explaining Customers a Machine-Learning Model

Benjamin Adrian[1], Markus Ebbecke[1], and Sebastian Ebert[1,2]

[1] Insiders Technologies GmbH, Kaiserslautern, Germany
{B.Adrian, M.Ebbecke, S.Ebert}@insiders-technologies.de
[2] Department of Computer Science, University of Kaiserslautern, Kaiserslautern, Germany

Abstract. When explaining customers that the artificial intelligence approach of our products automatically adapts document classifiers on training documents by applying statistical machine-learning, their reaction is similar like if we would tell them about an artificial intelligence in car breaks. Most likely they would dislike it, because they want full control on their data processors. Hence, we sell the Auto Classifier approach, which is the transparent and explainable extension of the respective machine learning components. This demo description presents this approach of providing customers full controls over document classifiers, which is part of nearly all products within Insiders Technologies' product line.

Keywords. Machine Learning, Explanation

1 Introduction

"Customers want maximum control. However, they also demand high degrees of automation." When developing, customizing, and selling our products, we had to learn this lesson, by hard. Black box machine-learning models, such as artificial neural networks will not be accepted by any customer, who processes sensitive data. Customers are willing to accept errors in classifications, if the classification system provides transparent and therefore understandable explanations [1]. However, explaining customers the need for collecting a preferable large set of training documents for each class is not a trivial task. At latest, when sufficient classification ratios cannot be achieved, because of bad training data, responsible consultants require a solid basis for discussion and explanation.

Hence, we developed an interaction and visualization framework for native document classifiers, the Auto Classifier [2]. It reveals and explains each phase of the machine learning process, including, sensing, segmentation, feature extraction, classification and post processing [3].

S. Wölfl (Ed.): Poster and Demo Track of the 35th German Conference on
Artificial Intelligence (KI-2012), pp. 115-119, 2012.

2 Machine-Learning Components in Document Classification

The main contribution of the Auto Classifier is to support users in required process steps for training and running a document classifier [2]. Every day, customers such as insurance companies receive a large number of documents at their document entry point. By using a categorization scheme, customers classify these documents, i.e., as invoice, damage report, complaint. It is not unusual that customers maintain large categorization hierarchies consisting of more than 200 categories. The document classification assigns an incoming document to one or several predefined categories. For reaching this goal, each category is described by a preferably large set of example documents. Therefore, a training algorithm computes statistics on discriminative features of each category. These statistics can be created by models such as Naïve Bayes [3], Support Vector Machines [4], or Logistic Regression [3]. The general process for creating a classifier is described as follows [3]:

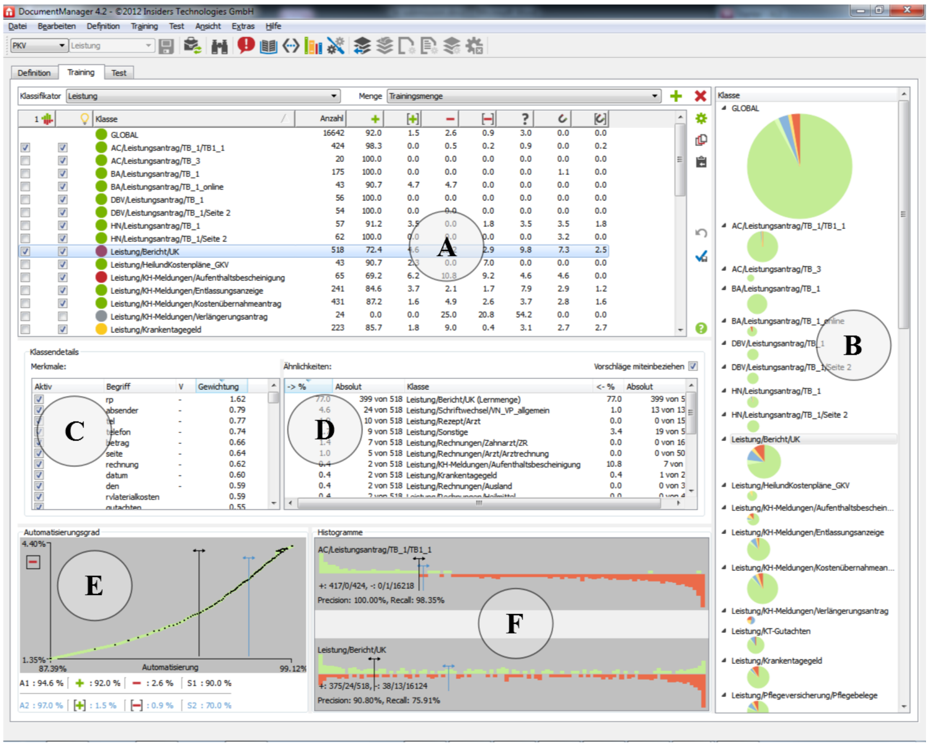
Sensing: Incoming documents have to be converted to a standard format. In case of paper based documents, these papers have to be scanned and processed by an optical character recognition (OCR). The document sensing results in a representation form providing information about contained text and layout. Some documents are of low OCR quality. Hence, they are not suitable as training examples and would be misclassified, later on. Customers should understand such problem documents and use the Auto Classifier to ignore documents with a limited number of recognized characters.

Segmentation: Either resulting from OCR processing, or resulting from performing white-space-based word segmentation, each document is represented as a bag of words. The Auto Classifier enables users to inspect and visualize this representation in terms of documents and finally of classes.

Feature extraction: By using term-based metrics such as term frequency (TF), or the inverse document frequency (IDF), words can be rated with respect to their distribution in training data. Words of high frequency can also be removed by performing a stop word removal since they do not contain any class specific information. The Auto Classifier visualizes the word rating and clarifies removed words. After modifying the feature description, users can easily retrain and re-evaluate their classifiers for inspecting the impact of these modifications on the accuracy of classification results.

Classification: Following [5], we train a linear classifier from the LIBLINEAR [4] library. Here, we perform a grid search for finding optimal values for the penalty parameter C, which defines costs for misclassified examples in training data. The overall process is performed by an x-fold cross validation. P, recall, and F-measure [5] are used as metrics for rating a classifier's performance.

Depending on the size of training data the user can chose the number x of folds. He is also enabled to set the value of C by hand. The result of training and cross validation can be inspected and visualized by using confusion matrices or several forms of diagrams (see Section 3).


DocumentManager 4.2 - ©2012 Insiders Technologies GmbH
Datei
Bearbeiten
Definition
Training
Test
Ansicht
Extras
Hilfe
PKV
Leistung
Definition
Training
Test
A
B
C
D
E
F
Klasse
GLOBAL
AC/Leistungsantrag/TB_1/TB1_1
AC/Leistungsantrag/TB_3
BA/Leistungsantrag/TB_1
BA/Leistungsantrag/TB_1_online
DBV/Leistungsantrag/TB_1
DBV/Leistungsantrag/TB_1/Seite 2
HN/Leistungsantrag/TB_1
HN/Leistungsantrag/TB_1/Seite 2
Leistung/Bericht/UK
Leistung/HeilundKostenpläne_GKV
Leistung/KH-Meldungen/Aufenthaltsbeschein...
Leistung/KH-Meldungen/Entlassungsanzeige
Leistung/KH-Meldungen/Kostenübernahmean...
Leistung/KH-Meldungen/Verlängerungsantrag
Leistung/KT-Gutachten
Leistung/Krankentagegeld
Leistung/Pflegeversicherung/Pflegebelege
Automatisierungsgrad
4.40%
1.35%
87.39%
Automatisierung
99.12%
Histogramme
AC/Leistungsantrag/TB_1/TB1_1
Leistung/Bericht/UK

Klassifikator: Leistung
Menge: Trainingsmenge

1		Klasse	Anzahl	+	[+]	−	[−]	?	↻	[↻]
		● GLOBAL	16642	92.0	1.5	2.6	0.9	3.0	0.0	0.0
☑	☑	● AC/Leistungsantrag/TB_1/TB1_1	424	98.3	0.0	0.5	0.2	0.9	0.0	0.2
□	☑	● AC/Leistungsantrag/TB_3	20	100.0	0.0	0.0	0.0	0.0	0.0	0.0
□	☑	● BA/Leistungsantrag/TB_1	175	100.0	0.0	0.0	0.0	0.0	1.1	0.0
□	☑	● BA/Leistungsantrag/TB_1_online	43	90.7	4.7	4.7	0.0	0.0	0.0	0.0
□	☑	● DBV/Leistungsantrag/TB_1	56	100.0	0.0	0.0	0.0	0.0	0.0	0.0
□	☑	● DBV/Leistungsantrag/TB_1/Seite 2	54	100.0	0.0	0.0	0.0	0.0	0.0	0.0
□	☑	● HN/Leistungsantrag/TB_1	57	91.2	3.5	0.0	1.8	3.5	3.5	1.8
□	☑	● HN/Leistungsantrag/TB_1/Seite 2	62	100.0	0.0	0.0	0.0	0.0	3.2	0.0
☑	☑	● Leistung/Bericht/UK	518	72.4	4.6	[illegible]	2.9	9.8	7.3	2.5
□	☑	● Leistung/HeilundKostenpläne_GKV	43	90.7	2.3	0.0	7.0	0.0	0.0	0.0
□	☑	● Leistung/KH-Meldungen/Aufenthaltsbescheinigung	65	69.2	6.2	10.8	9.2	4.6	4.6	0.0
□	☑	● Leistung/KH-Meldungen/Entlassungsanzeige	241	84.6	3.7	2.1	1.7	7.9	2.9	1.2
□	☑	● Leistung/KH-Meldungen/Kostenübernahmeantrag	431	87.2	1.6	4.9	2.6	3.7	2.8	1.6
□	□	● Leistung/KH-Meldungen/Verlängerungsantrag	24	0.0	0.0	25.0	20.8	54.2	0.0	0.0
□	☑	● Leistung/Krankentagegeld	223	85.7	1.8	9.0	0.4	3.1	2.7	2.7

Klassendetails

Merkmale:

Aktiv	Begriff	V	Gewichtung
☑	rp	-	1.62
☑	absender	-	0.79
☑	tel	-	0.77
☑	telefon	-	0.74
☑	betrag	-	0.66
☑	seite	-	0.64
☑	rechnung	-	0.62
☑	datum	-	0.60
☑	den	-	0.59
☑	rvlaterialkosten		0.59
☑	gutachten		0.55

Ähnlichkeiten: Vorschläge miteinbeziehen ☑

-> %	Absolut	Klasse	<- %	Absolut
77.0	399 von 518	Leistung/Bericht/UK (Lernmenge)	77.0	399 von 5
4.6	24 von 518	Leistung/Schriftwechsel/VN_VP_allgemein	1.0	13 von 13
1.9	10 von 518	Leistung/Rezept/Arzt	0.0	0 von 15
1.7	9 von 518	Leistung/Sonstige	3.4	19 von 5
1.4	7 von 518	Leistung/Rechnungen/Zahnarzt/ZR	0.0	0 von 16
1.0	5 von 518	Leistung/Rechnungen/Arzt/Arztrechnung	0.0	0 von 50
0.4	2 von 518	Leistung/KH-Meldungen/Aufenthaltsbescheinigung	10.8	7 von
0.4	2 von 518	Leistung/Krankentagegeld	0.4	1 von 2
0.4	2 von 518	Leistung/Rechnungen/Ausland	0.0	0 von 3
0.4	2 von 518	Leistung/Rechnungen/Heilmittel	0.0	0 von 4

A1 : 94.6 % | + : 92.0 % | − : 2.6 % | S1 : 90.0 %

A2 : 97.0 % | [+] : 1.5 % | [−] : 0.9 % | S2 : 70.0 %

AC/Leistungsantrag/TB_1/TB1_1: +: 417/0/424, -: 0/1/16218
Precision: 100.00%, Recall: 98.35%

Leistung/Bericht/UK: +: 375/24/518, -: 38/13/16124
Precision: 90.80%, Recall: 75.91%

Fig. 1. Screenshot of the Auto Classifier Training UI

Post processing: After training classifiers, users can inspect the representation of each class by looking at the most important words. Here, the Auto Classifier allows a manual filtering of words (e.g., Remove a person's name if the original training documents are all e-mails from a single customer employee.).

The Auto Classifier allows users inspecting the performance of each classifier. Because Logistic Regression [4] provides certainty values for each classified example, we decided to use this classification model. This enables users to define accuracy thresholds within a range of 0.0 and 1.0 on certainties. We defined these ranges as certain, likely, unsure, misclassified. The following colors are used referring to these ranges: green, blue, yellow, red.

3 Auto Classifier Training User Interface

The screenshot of the Auto Classifier training user interface (UI) in **Fig. 1** shows how we implemented the interaction features that were mentioned in the last section. The UI allows choosing between different classifiers and training and test sets. It consists of the following components:

- **(A) Classification area:** The table on the top side of the UI provides an overview of a classifier's performance. It lists containing classes of the training set. Relating to the selected test set, the numbers of examples as well as the percentage of correctly and incorrectly classified examples are listed. Dependent on given certainty thresholds, a class is marked as certain, likely, and unsure.
- **(B) Pie charts:** The pie charts on the right side of the UI visualize the distribution of certain, likely, unsure, and misclassified training examples.
- **Class details:** Beneath the classification area, details and properties of selected classes as well as distinctions to other classes are given:
  - **(C) Feature distribution:** This list on the left figures out importance of word features of a certain class, including their class weights.
  - **(D) Confusion matrix:** This table on the right is a serialized confusion matrix. It provides details on false positives and false negatives and therefore shows potentially overlapping classes.
- **(E) Degree of automation:** On the bottom left side of the UI, the user can define the three ranges of certain, likely, and unsure classifications on a function of classification errors.
- **(F) Histograms:** As an alternative visualization form, the histogram also allows users to define the three ranges of certain, likely, and unsure classifications directly on the distribution of training examples on bins of classification certainties.

4 Demo outline

The presentation of the Auto Classifier is going to be a live system demonstration on several datasets with an existing categorization scheme. Visitors are invited to using the Auto Classifier for training, modifying the feature space or just exploring different threshold ratios. The live demo of the user interface provides insights on end customers' expectations and needs when integrating a machine-learning system into their business processes.

5 Related Work

Compared to existing machine-learning toolkits such as WEKA [6], Rapid Miner [7], or KNIME [8], the Auto Classifiers' focus is not set on providing various data analyses and processing techniques. Instead, it offers a single way for solving a document categorization problem. For this way, the Auto Classifier provides plenty of transparent user interactions and data visualizations. Finally, it may not offer the best technology for solving a problem up to 99 % precision and recall. But it explains users the shape of their data and provides details and interactions on each single classification item. This enables users exploring the best solution for a given corpus of documents and set of categories.

6 Conclusion

The Auto Classifier is a key component of Insiders Technologies' product portfolio [2]. We showed how the UI can support the user in understanding classification internals, such that he is able to modify the results according to his needs. The ability to figure out, explain, and interact with classification internals was often the determining argument for customers to buy our product.

Acknowledgements

This work was funded by the German Federal Ministry of Education and Research (BMBF) in the INDINET project under grant number 01IC10S04I.

References

1. Forcher, B., Agne, S., Dengel, A., Gillmann, M., Roth-Berghofer, T.: Towards understandable explanations for Document Analysis Systems. 10th IAPR International Workshop on Document Analysis Systems (DAS 2012), Surfers Paradise, Queensland, Australia, (2012)
2. Klein, B., Dengel, A., Fordan, A.: smartFIX: An Adaptive System for Document Analysis and Understanding. In: Reading and Learning-Adaptive Content Recognition, LNCS 2956, Springer Verlag, (2004)
3. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern classification. Wiley (2001)
4. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J:. LIBLINEAR: A library for large linear classification, *Journal of Machine Learning Research* 9, 1871-1874, (2008)
5. Dumais, S., Platt, J., Heckerman, D., Sahami, M.: Inductive learning algorithms and representations for text categorization, Proceedings of the 7th int. conference on Information and knowledge management, p.148-155, Bethesda, Maryland, United States (1998)
6. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H: The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1 (2009).
7. Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., and Euler, T.: YALE: Rapid Prototyping for Complex Data Mining Tasks, in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06), (2006).
8. Silipo, R., Mazanetz, M.P.: The KNIME Cookbook, KNIME Press, Zürich, Switzerland, (2012)