Generating Explanations for Support Vector Machine Classification Results

Insu Song and Joachim Diederich

School of Business and IT
James Cook University Australia

Abstract—An explanation capability is crucial in security-sensitive domains such as medical applications. Although support vector machines (SVMs) have shown superior performance in a range of classification and regression tasks, SVMs like artificial neural networks (ANNs) lack an explanatory capability. There is a significant literature on obtaining human-comprehensible rules from SVMs and ANNs in order to explain “how” a decision was made or “why” a certain result was achieved, and this paper proposes a novel approach for SVM classifiers. The experiments describe a first attempt to generating explanations to “why” certain text contents are depressing or funny using a decompositional approach. Learned model parameters are analysed to select informative feature components and apply filtering to select subsets of more relevant components.

B. Background

This work is based on text mining and in particular text classification. The following section provides a brief overview of the core techniques, focussing on support vector machines (SVMs), the significance of generating human-comprehensible explanations from SVMs, and what it means to explain the decision-making process of a machine learning system to a human user who may not be an expert on artificial intelligence.

1) Support Vector Machines: Cortes & Vapnik [5] introduce support vector machines which are a novel approach to machine learning. SVMs are based on the structural risk minimization principle in order to overcome the overfitting problem. Support vector machines find the hypotheses out of the hypothesis space $H$ of a learning system which approximately minimises the bound on the actual error by controlling the empirical error using training samples and the complexity of the model using the VC-dimension of $H$. SVMs are very universal learning systems [6]. In their basic form, SVMs learn maximal margin hyperplanes (linear threshold functions). A hyperplane can be defined by a weight vector $w$ and a bias $b$:

$$w \cdot x + b = 0$$

The corresponding threshold function for an input vector $x$ is then given by:

$$f(x) = \text{sign}(w \cdot x + b)$$

However, it is possible learn polynomial classifiers, radial basis function (RBF) networks and three or more layered neural networks by mapping input data $x$ to some other (possibly infinite dimensional) feature space $\phi(x)$ and using kernel functions $K(x_i, x_j)$ to obtain dot products, $\phi(x_i) \cdot \phi(x_j)$, of feature data.

2) Explanations - The foundation: To illustrate why it is significant to provide explanation capability to SVMs, let us consider the case where a medical doctor tells a patient of a diagnosis that whether the result is positive or not. It is a norm that a doctor usually also includes explanations of a diagnosis result. For instance, the explanations may be a deductive argument which comprises a list of observed symptoms of the patient, a list of possible causes, and modus ponens (the rule of inference) for deriving the conclusion.

Thagard and Litt [7] illustrate several major approaches to generating explanations. The classical view is that explanation
is a deductive argument including background knowledge and inference rules such as modus ponens. The inference rules allow the sequential application of “if-then-else” statements in order to justify an explanatory target. Whenever no precise knowledge is available, explanatory schemas or probabilistic rules can be used.

Cawsey [8] is using a very simple definition of explanation: “In general an explanation is something which makes some piece of knowledge clear to the hearer. . . . The explanation is complete when the hearer is satisfied with the reply and understands the piece of knowledge” [8]. Hence, explanation is based on an “information need”.

3) Generating Explanations from SVMs: Much of works trying to provide an explanation capability to SVMs have been on rule extraction [4] following the footsteps of the earlier effort to obtain human-comprehensible rules from artificial neural networks (ANNs). One approach of classifying rule-extraction is the transclucency dimension of classification, which comprises of the decompositional and pedagogical (or learning based) techniques [9].

The decompositional approach relies on the degree to which the internal representation of the ANN is accessible to the rule extraction technique. The basic strategy of decompositional techniques is to extract rules at the level of each individual hidden and output unit within the trained ANN. In general, decompositional rule extraction techniques incorporate some form of analysis of the weight vector and associated bias (threshold) of each unit in the trained ANN. Then, by treating each unit in the ANN as an isolated entity, decompositional techniques initially generate rules in which the antecedents and consequents are expressed in terms which are local to the unit from which they are derived.

In contrast to the decompositional approaches, the strategy of the pedagogical approaches is to view the trained ANN at the minimum possible level of granularity i.e. as a single entity or alternatively as a “black box”. The focus is on finding rules that map the ANN inputs (i.e. the attribute/value pairs from the problem domain) directly to outputs [10]. In addition to these two main categories, Andrews et al. [9] also proposed a third category which they labelled as “eclectic” to accommodate those rule extraction techniques which incorporate elements of both the decompositional and pedagogical approaches.

C. Overview

The reminder of this paper summarises experiments and their results: text classification, explanation generation for classification results, and statistical analysis on the model parameters that are generated for depression poems.

II. EXPERIMENTAL EVALUATION

A. Methodology

A preliminary study has been undertaken on generating explanations of depression poems and classifications of online text messages. The depression poems were obtained from the Internet (PoetryAmerica.com) and comprise of a total of 76 poems: 56 depression poems and 20 funny poems. The online text messages were obtained from Usenet news groups and comprise of a total of 350 sentences: 297 open questions and 53 closed questions. The resulting text documents are represented as attribute-value vectors (“bag of words” representation) where each distinct word corresponds to a feature whose value is the frequency of the word in the text sample. Values were transformed with regard to the length of the sample. For the poem data set, functional words were removed and stemming were performed on each extracted text. For the question data set, all words were used. In summary, input vectors for machine learning consist of attributes (the words used in the sample) and values (the transformed frequency of the words). Outputs are depression versus funny and open question versus closed question, that is, binary decision tasks were learned. Clearly, the expressive power of the resulting explanations is limited by this “bag of words” representation.

For LOO (leave-one-out) cross validation, 76 and 350 SVM models were generated using the linear kernel for the poem and online message text data sets, respectively. Thus, each model is used to classify one document. An SVM model is defined by support vectors $x_i$ and associated parameters. The decision value of a text sample (represented as a feature vector $x$) is then obtained as follows:

$$d(x) = w \cdot \phi(x) + b = \sum_{i \in SV} \alpha_i y_i K(x_i, x) + b$$

where $x_i$ are support vectors and $x$ is the feature vector, $\alpha_i$ are Lagrangian multipliers, and $b$ is the offset. The antecedent of the rule of inference is then $d(x) > 0$: if $d(x) > 0$, the feature vector $x$ is positive or else negative. We use this insight into the SVM models to define three types of explanations:

1) Explanation A comprising of all the terms contributing to the decision value $d(x)$.
2) Explanation B comprising of the top-contributing terms that are sufficient to classify the features.
3) Explanation C comprising of the top-contributing terms that also have their sensitivity values $\partial d/\partial x_j$ greater than a set threshold value $\tau$.

Technical details on generating each explanation types are described in Section III. This approach is clearly a decompositional approach: analysis on the model parameters to select informative components and selecting subsets of more relevant components. Figure 1 summaries the significance of each types of explanations. It plots sensitivity, contribution, and word rank of all feature components of the depression poem text data set. It shows that sample feature components having higher ranking order (more frequent words) and higher sensitivity values tend to have larger contribution values. This suggests that feature components having higher sensitivity values and higher ranking order provide greater information in decision making than other feature components. It also shows that most of large contributions are made by more frequent words (high rank words).

B. Result

Support vector machines trained on the poem and online text message data sets achieved accuracy of 94% and 98%,
respectively. The sensitivity values are adjusted manually to obtain reasonable amount of terms for Explanation type C. Sample explanations of a depression poem are provided below (Explanation A samples are too big to show here):

1) Explanation B: small 6, just 0, cry 1, beware 1, hear 0, strong 0, worry 2, must 3, grow 2, feet 2, slowly 2
2) Explanation C: small 6, worry 2, must 3, grow 2, feet 2

The numbers indicate relative contribution values to the decision value \( d(x) \), Sensitivity-filtering (Explanation C) eliminates some of less sensitive terms (bold-faced terms) from Explanation B.

Sample explanations of a funny poem are provided below:

1) Explanation B: away -9, muscle 0, day -2, back -1, turn -5, fear -3, hand -2, came -2, dead -3, head -6, bite -2
2) Explanation C: away -9, turn -5, head -6

Negative cases have negative contribution values.

The explanations of the question data sets are much shorter. Explanations of an open question, “What is the dolphin species seen in most oceanaria?” is provided below:

1) Explanation A: species 6, most 3, what 107, in 1
2) Explanation B: what 107
3) Explanation C: what 107

Explanations of a closed question, “Do dolphins live shorter in captivity?” is provided below:

1) Explanation A: dolphins -31, live -14
2) Explanation B: dolphins -31
3) Explanation C: {empty}

As expected, questions are explained by the presence or absence of question words, such as what, why, and how.

III. GENERATING EXPLANATIONS FROM SVM MODELS

In order to calculate the contribution values of each feature components of a feature vector \( x \), we use the centroid \( C \) of the population, which is estimated using the centroid \( C_{sv} \) of the support vectors:

\[
C_{sv} = \frac{1}{N_{sv}} \sum_{i \in SV} \phi(x_i)
\]

where \( N_{sv} \) is the number of support vectors. We can then calculate the deviation of a feature vector \( x \) from the estimated population centroid:

\[
D(x) = \phi(x) - C_{sv}
\]

Suppose \( C_{sv} \) is on the hyperplane: \( w \cdot C_{sv} \approx -b \). Then, we can obtain the decision value \( d(x) \) using the deviation \( D(x) \):

\[
d(x) = w \cdot (D(x) + C_{sv}) + b \\
\approx w \cdot (\phi(x) - C_{sv})
\]

If \( K \) is the linear kernel, we can estimate the contribution of each \( j \)-th feature component \( x_j \) as follows:

\[
C_{sv,j} = \frac{1}{N_{sv}} \sum_{i \in SV} x_{i,j}
\]

\[
d(x)_j = \sum_{i \in SV} \alpha_i y_i x_{i,j} (x_j - C_{sv,j})
\]

Now, for a feature vector \( x \), we can explain why a sample is positive (negative) by listing the feature components that contribute to the decision value. That is, we can rank the feature components of a feature vector according to the amount of contributions made by the feature components. This is used as the basis of the explanation type A. We can also calculate the sum of all negative (positive) contributions and choose the top \( N \) positive (negative) contributions that are sufficient to push the decision value to positive (negative). This is used as the basis of the explanation type B.

A. Filtering Explanations with Sensitivity

Training a support vector machine for a data set of interest generates a hyperplane, which can be used to obtain the distance of a feature vector to the hyperplane to classify. The distance is normal to the hyperplane and thus the importance of a feature component can be measured as the rate of change
of the distance with respect to the feature component. This can be easily obtained for a linear classifier as follows:

$$\frac{\partial d(x)}{\partial x_j} = \sum_{i \in SV} \alpha_i y_i x_{i,j}$$

where $d(x)$ is the decision value which is proportional to the distance of feature $x$ to the hyperplane, $x_j$ is the $j$-th component of the feature $x$, and $x_{i,j}$ is the $j$-th component of a support vector $x_i$. As we can see from the above equation, the importance of the $j$-th component for the hyperplane is the sum of $j$-th component of the support vectors multiplied by the class label and the Lagrange multipliers. Figure 2 shows a histogram of sensitivity values of feature elements for the poem text data set. Greater population is centred at sensitivity value 0. This suggests that feature components having higher sensitivity values will provide more information on the decisions. This is also suggested in Figure 1 that most of the contributions are made by terms with higher sensitivity values.

Figure 3 shows the relationship between word ranks and sensitivity values of sample feature components. Feature components with higher sensitivity values tend to have higher ranking order (i.e., lower rank value). The relationship between the word rank and sensitivity for the poem text data set can be summarised as follows:

$$\text{rank} \leq a \frac{1}{|\text{sensitivity}|}$$

for some positive constant $a > 0$. Similar relationships are also observed between the word rank and contribution across different text data sets ($b > 0$):

$$\text{rank} \leq b \frac{1}{|\text{contribution}|}$$

However contribution values are proportional to sensitivity values ($c > 0$):

$$|\text{contribution}| \leq c|\text{sensitivity}|$$

which strongly suggests the two-step filtering (Explanation type B and Explanation type C).

For non-linear cases, we calculate partial derivatives of the kernels. As an example, let us consider the polynomial kernel:

$$K(x_i \cdot x)_{\gamma,d} = (\gamma x_i \cdot x + r)^d.$$

$$\frac{\partial d(x)}{\partial x_j} = \sum_{i \in SV} \alpha_i y_i d_j \gamma x_{i,j} (\gamma x_i \cdot x + r)^d$$

$$= d \gamma \sum_{i \in SV} \alpha_i y_i x_{i,j} K(x_i \cdot x)_{\gamma,d-1}$$

This shows that the importance of the $j$-th feature component for the hyperplane is a combination of other feature components weighted by support vector components. In this case, one way of obtaining sensitivity values would be averaging the sensitivity values over a range of selected points, such as support vectors.

IV. DISCUSSIONS AND FUTURE WORK

This is the first report of a novel approach to explaining psychiatric classification, in particular explaining depression poem text contents. Considerable further research is required. The approach of extracting some piece of knowledge using machine learning in explaining psychiatric assessments has the potential to improve the usability of machine learning techniques in the medical and security domains. This approach of extracting explanations using some form of analysis on machine learning and associated parameters can be further expanded by using alternative feature representations of text data sets, such as concept terms or semantic terms. There is massive potential of incorporating these sophisticated information extraction technologies within psychiatry and in medicine more generally.

REFERENCES