4F-6

ユーザ入力における適正入力支援技術の提案

藤野 友也　　平井 規郎　　石井 篤
三菱電機株式会社 情報技術総合研究所

1　はじめに

データ分析の重要性は年々高まってきているが、ユーザ入力により蓄積されたデータには、誤入力などにより本来の値と異なる値が含まれている場合が多い。そのようなデータを分析する場合、前処理としてデータクレンジングが必要になるが、要求されるコストは一般に大きい。入力から時間が経過するにつれて、本来の値を確認するためのコストは増大する。

文字列の入力に関しては、入力途中で後続の文字列を補完する機能[1]など、入力負荷の低減および誤入力の抑制を提供する環境がポータルサイトなどに整備されてきているが、数値入力に対する支援環境はいまだ整備されていない。

本論文では、次の方針に基づく適正入力支援技術を提案する。入力段階で、入力済みの項目と過去の入力履歴から、新たに入力しようとしている値をその場で評価する。その結果、入力値が妥当でないと判断した場合には、そのレベルに応じた警告を適切に与えることで、注意を喚起し誤入力を防止する。

ここでは、妥当性の評価に、主成分分析により得られた特徴空間と入力値との距離を用いる方法を提案する。

2　適正入力支援

ユーザ入力時に混入する誤値の原因としては、大きく以下の 3 つが挙げられる。

(A) 入力ミス
(B) 情報誤認
(C) 人為的調整

「入力ミス」は、ユーザが意図した値と異なる値を入力することで、主にタイプミスである。「情報誤認」は、目盛の読み違いなどで、誤った値を正しい値と誤認して入力することである。「人為的調整」は、入力値が入力者の能力評価などに関係する場合、評価が高くなるよう調整して入力することである。

Advanced interface to suggest user's correction for an improper input.
Tomoya FUJINO, Norio HIRAI, Atsushi ISHII,
Mitsubishi Electric Corporation,
Information Technology R&D Center

入力時点で警告がされる場合、(A) は直ちに訂正できる。(B) は、情報源を再確認することで、後に発見･修正する場合と比べて低コストで訂正できる。(C) に対しては、故意の誤入力への警告により、心理的な抑止効果が期待できるが、本論文ではその効果の評価は行わない。

混入した誤値は、本来のデータが持つ特性と関連の薄い値になるため、本来の特性と比較することで検出できることがある。当然、本来の値が特性に従わない場合もあるが、その場合は、その値が重要な意味を持つことが多い。データ本来の特性から外れた値を統計的に検出し、ユーザに認知させることで、データ品質の向上や知見獲得を支援する結果となるものと考える。

ユーザへの認知方法としては、表 1のように、入力値の妥当性に応じて警告のレベルを変更することで、ユーザが警告に麻痺することを回避し、的確な警告を与える方法をとる。妥当性無の判定は、データの入力規則との照合によってのみ行う。統計的な判定は、妥当性高～低の範囲の判断のみとし、操作の阻害を極力抑制する。

表 1　妥当性ごとの警告方針

妥当性	方針	表現
高	表示上の変化なし	通常色を維持
中	操作を阻害せず、表示の変更のみ	警戒色への変更
低		ポップアップ
無	操作の中断	ダイアログ表示

なお、妥当な値をシステムが提示することは、データ品質向上の目的には逆効果と考えられる。情報源への再確認を促すことが重要である。

3　主成分分析による妥当性評価

項目 X へ入力中の値 v について、その妥当性を評価する素直な方法は、項目 X の過去の値を確定済みの条件 C で絞込んだ後の分布から判断することである。しかし、データ量が大きい場合、妥当性判定の度に時間を要する絞込みが必要となるため、即応性の確保が困難である。

主成分分析を用いることによって、データ空間におけるデータの分布を効率よく線形近似する、より次元数の小さい空間（特徴空間）を取得することができる。特徴空間が得られた場合、項目 X を含む新たなデータ標本に対して、特徴

空間との距離が最小となるように項目 X の値を決定することで、入力されるべき値 $p(C)$ を推定することが可能である。

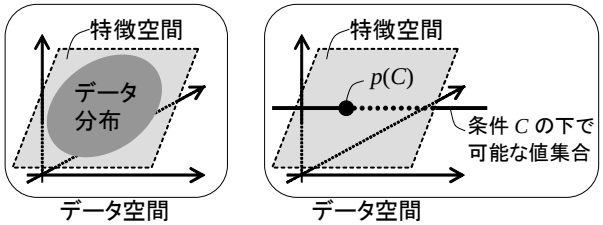

図 1 特徴空間と推定値 $p(C)$

特徴空間の算出には時間を要するが、特徴空間から $p(C)$ を得る演算は実時間で可能である。

ここで入力中の値 v の妥当性を、次のように定義する。条件 C のもとでの、項目 X の標準偏差を $\sigma(C)$ とするとき、入力中の値の妥当性評価値 $A(v,C)$ を、正規分布になぞらえて次式で示す。

$$A(v,C)=\exp\left(-\frac{\{v-p(C)\}^2}{2\sigma(C)^2}\right)$$

ここでは、正規分布における σ 範囲内に対応する $A(v,C)\geq 0.6$ の場合の妥当性を「高」、3σ 範囲外に対応する $A(v,C)\leq 0.01$ の場合の妥当性を「低」、それ以外の場合の妥当性を「中」とする。

4 実装例

アメダス測定データの入力に対する実装例を紹介する。アメダス観測データは、観測所名, 年, 月, 日, 時,気温, 降水量, 日照時間, 風速および風向により構成される。これらの値は本来自動的に測定され蓄積されるが、ここでは仮に、手作業で入力する場合を考える。妥当性評価に用いる特徴空間は、主要都市 100 地点で、1976 年から 2003 年に時別測定された、2450 万件のアメダス観測データから、場所･月ごとに用意した。

例えば、観測地を札幌、月を 4 月とした場合、気温として「30」を入力すると、妥当性が低と判定され、ポップアップで警告が表示される。ユーザは、この警告を無視しても良い。(札幌 4 月の気温の平均値は 6.84、標準偏差は 4.35)

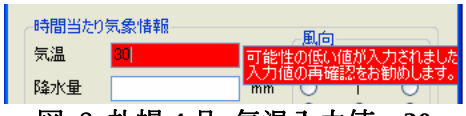

図 2 札幌 4 月 気温入力値：30

気温を「20」へ変更すると、ポップアップは消えるが、入力欄は濃い赤のままであり、まだ特性から外れていることを入力者に認識させる。

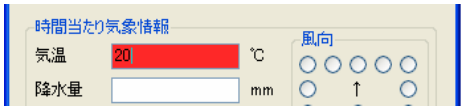

図 3 札幌 4 月 気温入力値：20

観測地を那覇へ変更して同じ値を入力すると、下図の通り妥当な値であると判定される。(那覇 4 月の気温の平均値は 21.4、標準偏差は 2.72)

図 4 那覇 4 月 気温入力値：20

5 精度評価

特徴空間による妥当性評価手法の精度評価を行った。評価には、2005 年 7 月から 2006 年 6 月までの主要 6 都市のアメダス時別データを用いた。センサによる観測値(正常値)と、人工的に入力誤り (隣接キー誤打、隣接キー同時押下、文字順序逆転、文字欠落) をランダムに付加した値とで、妥当性が「低」と判定される割合を比較した。隣接キーはテンキーのものとした。

表 2 妥当性が「低」と判定された割合

観測項目	正常値	誤入力値
気温	0. 020	0. 387
降水量	0. 391	0. 886
日照時間	0. 115	0. 791
平均風速	0. 003	0. 373

表 2より、例えば気温に関して、正常値が入力された場合に警告が発生する割合は 2%、誤入力値が入力された場合は 38. 7%であり、誤入力に対する、より高い警告発生確率が確認できる。入力誤りを効果的に検知し警告することで、ユーザに誤入力を認識させることができる。

6 結論と今後の課題

本論文では、データ入力時に、過去の履歴から入力値の妥当性を判断し、妥当でないと判定される値に対し警告を与える、入力インタフェースを利用した適正入力支援技術を提案した。本技術により、入力誤りの早期発見と、それに伴う低コストでの対策が実現できる。

今後は、ベイジアンネットワークなど、より柔軟な妥当性判定手法による評価を行っていく。

参考文献

[1] Bast H. and Weber I, “Type Less, Find More: Fast Autocompletion Search with a Succinct Index,” In *Proceedings of SIGIR '06*, ACM Press, New York, 364-371, 2006.